

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

SIGHT LINES

A ROOM-TEMPERATURE DETECTOR FOR
LONG-WAVELENGTH INFRARED RADIATION **PAGE 85**

IMMUNOLOGY

DIGGING THE DIRT

How squeaky clean mice
may be ruining research

PAGE 16

CONDENSED-MATTER PHYSICS

GRAPHENE IN A TWIST

Rotated sheets create
unusual superconductor

PAGES 37, 43 & 80

MICROBIOLOGY

PERSISTENT PROBLEM

New antibiotics tackle
dormant bacteria

PAGES 40 & 103

[NATURE.COM/NATURE](https://www.nature.com/nature)

5 April 2018

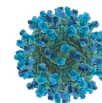
Vol. 556, No. 7699

THIS WEEK

EDITORIALS

WORLD VIEW Data companies must address the techlash **p.7**

FLASH Brief signal of shortest ever supernova **p.8**



DISEASE RNA remedy beats West Nile virus **p.9**

Nature: the truth

Myths always circulate about Nature's editorial processes and policies. Here is an attempt to dispel them.

Myth 1: Perhaps the longest-lived myth is that publishing a preprint of a paper submitted to this journal will pre-empt its consideration. Not true, as has been said in these columns before. For more than 20 years, we have had a policy of treating preprints as equivalent to conference talks: intra-researcher communication that encourages informal feedback and leads to better papers.

Myth 2: Nature journals do not want senior researchers to involve junior colleagues in the confidential process of peer review. Untrue. We positively encourage such involvement, to help graduate students and postdocs gain experience with due oversight. We ask that they be identified, give them credit and may well go to them directly for advice on subsequent papers.

Myth 3: Referees can veto papers. Only on technical grounds. It has always been the editors who select which papers *Nature* publishes, even though referees' assessments of significance are influential. We always heed technical comments, but reserve the right to disagree with a referee's recommendation as to whether publication is warranted.

Myth 4: The authorship of a paper — including country and institution — influences *Nature's* decision on whether to referee or publish it. Untrue. We frequently publish papers from first-time authors, and frequently reject papers by highly reputable researchers on purely editorial grounds of the paper's significance. We offer the option of double-blind peer review for those authors who want it. We recognize the possibility of unconscious bias.

Myth 5: *Nature* editors choose papers for anticipated media coverage or citations. Not true. Assessment of significance is what counts. In many areas of research, citations do indeed reflect significance, and

we value such achievements. But many papers that we publish neither achieve nor are expected to achieve high numbers of citations. We value them, nevertheless, because we judge them to have intrinsic interest, or because of their potentially substantive impact on society.

Myth 6: *Nature* editors sometimes reject papers without reading them fully. Untrue.

Myth 7: Authors must prepare submitted manuscripts in the form consistent with our highly *Nature*-specific format guidelines. Not true. For submission purposes, we care only that the paper conforms roughly to our length stipulations, and that editors and referees can understand the claims and their bases. Figures and their legends do not have to be placed at the end of the text at submission stage. Only moving towards publication does the formatting matter.

Myth 8: Within the *Nature* journals system, in which authors may be offered a transfer of a rejected paper to another journal, the transferred paper may be underestimated by the receiving editors. Untrue. Editors assess papers on their own terms. Because such transfers are informed by our knowledge of our journals' criteria, one would expect a substantially lower rate of prompt editorial rejections and a higher rate of refereeing for such transfers than for direct submissions — which is indeed borne out by our statistics. For example, manuscripts transferred within the *Nature* family were sent to external reviewers in February 2018 twice as frequently as those submitted directly.

Myth 9: *Nature* editors never consider appeals. Not true.

Our Guide to Authors may not be sufficiently clear on some of these policies, and we are working to improve it. Above all, we hope that this Editorial will help researchers, and correct sometimes widespread misconceptions about *Nature's* processes and policies. ■

Cries for help

An outpouring on Twitter highlights the acute pressures on young scientists.

Poor mental health is an issue for many of our readers, as underscored by the response to a tweet sent by @NatureNews last week, which highlighted rates of depression and anxiety reported by postgraduate students (see go.nature.com/2gtjxq). The reaction blew us away: more than 1,900 retweets and around 230 replies.

"This is not one dimensional problem. Financial burden, hostile academia, red tape, tough job market, no proper career guidance. Take your pick," read one. "I'd love to see some of the comments under this thread published," wrote one responder. "There needs to be real conversation about this, not just observation." We

agree — which is why we are publishing some of the responses.

There is a problem with the culture in science, and it is one that loads an increasing burden on the shoulders of younger generations. The evidence suggests that they are feeling the effects. (Among the tweets, one proposed solution to improving the PhD: "treat it like professional training instead of indentured servitude with no hope of a career at the end?") It will take a while to change that culture — and, unfortunately, it will probably take almost as long for some in the community to realize the need for it to change. But change it must.

We intend to revisit this topic, starting in May, with a Careers Feature on depression. We want to hear more from readers on mental-health issues and the stresses that contribute. You can share your stories in confidence here: go.nature.com/stress-stories.

We thank those who have already told theirs. "I hold down three jobs to fund my PhD, living in hopes of funding, it's a constant strain," wrote one. "So many others out there like me, and sometimes I wonder if it's even worth it. The research community will lose so many great minds to issues like this. It needs to be changed." ■



Use our personal data for the common good

Technology giants should take lessons from the Human Genome Project and be data stewards, not data owners, says Hetan Shah.

Data science brings enormous potential for good — for example, to improve the delivery of public services, and even to track and fight modern slavery. No wonder researchers around the world — including members of my own organization, the Royal Statistical Society in London — have had their heads in their hands over headlines about how Facebook and the data-analytics company Cambridge Analytica might have handled personal data. We know that trustworthiness underpins public support for data innovation, and we have just seen what happens when that trust is lost.

Allegations that Cambridge Analytica obtained data on tens of millions of people from Facebook, in circumstances still being investigated by multiple regulators, and used them to target political advertising in the 2016 US presidential election has led to a 'techlash'. There's been a US\$60-billion fall in the value of the social-media giant, and a surge in people searching for how to delete their accounts.

Much behind the outcry has been hiding in plain sight. Too many data companies' business models are based on hovering up our personal data and selling them.

What can be done to restore trustworthiness? Social-media companies must do more than say sorry and vow to improve protections. They must adapt to ensure that data collected are used for the common good.

The techlash snarls together several concerns. One is the protection of privacy. Some have argued that this requires strengthening the ownership we have over our own data, allowing people to select or sell levels of data use. This is problematic: it assumes 'data about me' are data I own. But many personal data are created through interactions with other people or services — if I have a relationship with somebody, who owns those data? A better question is what right to privacy does each of us have? I also doubt that offering more options for ownership would bring much change. Relatively few people shop around for the cable-television company or energy provider with the best rates or services. Why might they be more active with data? Finally, such a data-ownership model would increase inequality. The well-off and the well-informed would be protected, leaving the vulnerable to trade their data away.

Smart privacy regulation is a better approach to curbing inappropriate use of personal data. The European Union is making strides with its new General Data Protection Regulation policy, which will come into force in May and give EU nations stronger powers to deal with data breaches. In this area, the United States could learn from Europe. More widely, information regulation and regulators around the world need strengthening. Many policies were set up when the collection and use of personal data were backwater issues. In particular, to create and implement the best policies, regulators must be able to pay competitive salaries to recruit

technical talent, or risk losing it to the very giants that need regulating.

There is more than privacy at stake. Facebook and other social-media companies are now information (and misinformation) providers that affect our democracies. They need to ensure that their algorithms do not promote misinformation as clickbait. It would be in Facebook's interest to nourish a system that creates reliable content to fuel its users' interactions. Facebook could do much good if it put just 1% of its profits into an independent trust to fund quality media, especially local media, and fact checkers.

Another issue for democracy is microtargeted political advertising. Claims that Cambridge Analytica made about its ability to use this tactic to change people's minds on political issues were probably overblown. Microtargeting is not inherently unethical, but it must be made fully transparent. We cannot do democracy in the dark.

There is also unease that technology companies will grow into unchecked data monopolies. It would be hard to break the companies up, because we would then lose the networked benefits we get from them as consumers. But how else might we ensure the use of data for the public good rather than for purely private gain?

Here are two proposals towards this goal.

First, governments should pass legislation to allow national statistical offices to gain anonymized access to large private-sector data sets under openly specified conditions. This provision was part of the United Kingdom's Digital Economy Act last year and will improve the ability of the UK Office for National Statistics to assess the economy and society for the public interest.

My second proposal is inspired by the legacy of John Sulston, who died last month. Sulston was known for his success in advocating for the Human Genome Project to be openly accessible to the science community, while a competitor sought to sequence the genome first and keep data proprietary.

Like Sulston, we should look for ways of making data available for the common interest. Intellectual-property rights expire after a fixed time period: what if, similarly, technology companies were allowed to use the data that they gather only for a limited period, say, five years? The data could then revert to a national charitable corporation that could provide access to certified researchers, who would both be held to account and be subject to scrutiny that ensure the data are used for the common good.

Technology companies would move from being data owners to becoming data stewards. ■

Hetan Shah is executive director of the Royal Statistical Society and visiting professor at the Policy Institute, King's College London.
Twitter: @HetanShah

**DATA COMPANIES
SHOULD WORK TO
BOOST
POTENTIAL
USE OF DATA
FOR THE
PUBLIC GOOD.**

SEVEN DAYS

The news in brief

TECHNOLOGY

AI investment

France will invest €1.5 billion (US\$1.8 billion) in artificial-intelligence (AI) research and innovation by 2022, as part of a national strategy unveiled by French President Emmanuel Macron at a conference in Paris on 29 March. INRIA, France's national computer-science agency, will coordinate the plan with other French research agencies and universities. The strategy aims to create networks of AI research institutes in four or five places across the country. At the conference, Google's London-based AI firm DeepMind also announced that it will establish a centre in Paris, its first in continental Europe. South Korean electronics giant Samsung said that it would create a large AI centre in or near the capital and employ about 100 researchers. Japanese firm Fujitsu announced plans to make France its European centre of AI research.

POLICY

Fuel efficiency

On 2 April, the US Environmental Protection Agency (EPA) announced plans to relax fuel-efficiency standards for vehicles manufactured between 2022 and 2025. The current standard, which was finalized in January 2017 after a review by former president Barack Obama's administration, would require the average fuel efficiency for new passenger cars and trucks to be 23.2 kilometres per litre by 2025 — 33% higher than the goal in 2010. EPA chief Scott Pruitt said that the standards had been set too high. He will also reconsider an EPA waiver that allows California to set fuel-efficiency standards independently of

the federal government, and which other states may choose to follow. The agency will work with the National Highway Traffic Safety Administration to establish a new standard.

CRISPR crops

The US Department of Agriculture (USDA) will not regulate plants produced using genome-editing techniques, including CRISPR-Cas9, the agency announced on 28 March — as long as they could also have been created using conventional breeding techniques. The engineered products cannot be plant pests or have been developed using plant pests, and they can't contain genes from distant species. Those plants will still be regulated. The policy reverses rules proposed under former US president Barack Obama. It's unclear whether

food produced from gene-edited plants will need to be labelled as such.

FUNDING

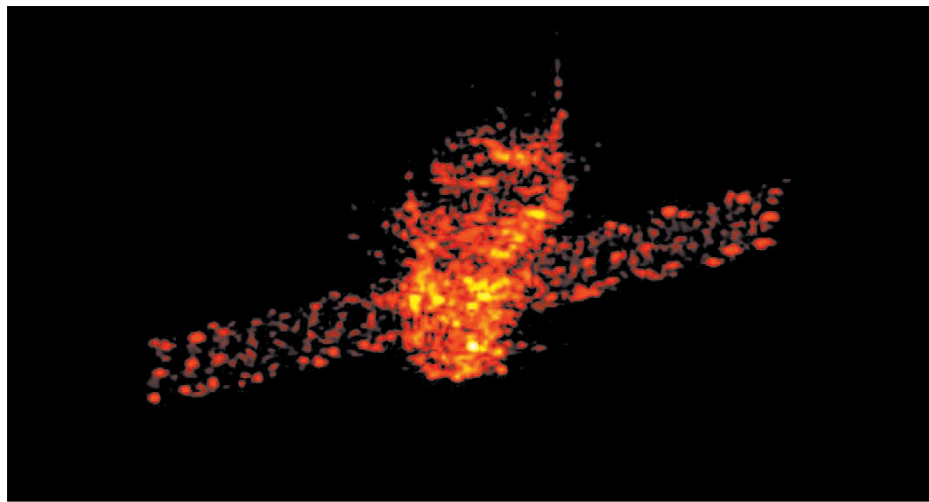
UK funding body

Britain's powerful new research-funding body officially came into existence on 1 April. UK Research and Innovation (UKRI), born out of sweeping higher-education and research reforms passed last year, brings together the country's seven research-funding councils and the business-focused Innovate UK. Research England, a new body that will oversee the research and knowledge-exchange activities of the now-defunct Higher Education Funding Council for England, will fall under the same umbrella. UKRI has an annual budget of £6 billion (US\$8.4 billion) and

is headed by Mark Walport, the government's former chief scientific adviser.

Singapore funding

Singapore's largest national research agency announced that groups doing basic scientific research will need to start competing for all their funding, as part of a reorganization unveiled on 27 March. The Agency for Science, Technology and Research (A*STAR) runs 18 research institutes; it will continue to guarantee core funding to those that partner with industry or focus on technology development. Basic scientists at those centres will be funded entirely through merit-based awards. The changes will take effect from 1 April, but will not affect projects that already have funding.



FRAUNHOFER/EPA/REX/SHUTTERSTOCK

Chinese space lab plummets to Earth

The defunct Chinese space station Tiangong-1 re-entered Earth's atmosphere on 2 April, breaking apart over the southern Pacific Ocean at around 00:15 UTC. The bus-sized spacecraft had been in orbit since 2011. Two groups of Chinese astronauts lived and worked

on it in 2012 and 2013, using it as a testbed for a future, larger space station. But in 2016, mission managers lost control of the spacecraft. As a result, it began an uncontrolled descent to Earth, although it was closely watched by organizations that track space debris.

NEWS IN FOCUS

BIOMEDICINE Cancer researchers push to relax criteria for inclusion in clinical trials **p.12**

ASTROPHYSICS Upgraded Italian detector spots signs of elusive dark matter **p.13**

PUBLISHING Potential changes to EU copyright rules rile researchers **p.14**



IMMUNOLOGY Could dirtier lab mice make for better science? **p.16**

DESIRÉE STOVER/NASA



The mirrors of the James Webb Space Telescope are designed to peer at the Universe's first stars.

ASTRONOMY

NASA reveals major delay for Hubble successor

James Webb Space Telescope woes could have broader effects on astrophysics programme.

BY ALEXANDRA WITZE

NASA will delay the launch of its ambitious James Webb Space Telescope (JWST) by nearly a year, until approximately May 2020. That is likely to push the cost of the mission — the most complex space-science telescope ever built — over the US\$8-billion limit set by the US Congress. It is the first major setback since NASA revised its plans for the project in 2011, after years of

slipping schedules and rising costs.

NASA announced the delay on 27 March, saying that engineers needed more time to assemble and test the components of the spacecraft at its main contractor, Northrop Grumman in Redondo Beach, California. Among other problems, the collapsible, tennis-court-sized sunshield that will protect the observatory's 6.5-metre mirror took weeks longer than expected to fold and refold during testing. "Frankly, the tests are taking longer to

complete than expected," Robert Lightfoot, NASA's acting administrator, told reporters.

The agency did not say how much the delay would cost, but some estimates suggest that it could add a few hundred million dollars to the project. In recent years, Congress has pushed NASA to hold down the cost of the telescope and other future missions.

The delay will affect NASA's astrophysics budget more broadly, including its next big planned space observatory, the Wide-Field ►

► Infrared Survey Telescope (WFIRST). The agency had aimed to spend more on WFIRST in the coming years as its contributions to the JWST shrank, trading off the end of one mission's development for the beginning of another. Now, the JWST delay risks compounding problems for WFIRST. Last month, US President Donald Trump proposed cancelling WFIRST; astronomers protested, and Congress gave the project \$150 million.

US astronomers ranked the JWST and WFIRST as the most important large space missions, respectively, in decadal surveys of their scientific priorities released in 2000 and 2010.

NASA delayed the JWST launch after an independent board of experts concluded this month that the project could not meet its June 2019 launch goal. Last September, the agency shifted to that target, abandoning the October 2018 launch date it had aimed for since rebooting the project in 2011.

Now, yet another independent panel — led by former aerospace executive Tom Young — will review the project's schedule. NASA will use these analyses to pick a more specific launch target in the coming months. “We have one shot to get this right before going into space,” says Thomas Zurbuchen, NASA's associate administrator for science.

Until recently, project managers were able to deal with schedule problems by moving work

tasks around locations. For instance, engineers decided to clean the telescope's mirror at NASA's Johnson Space Center in Houston before shipping it to the crowded spacecraft-assembly facility at Northrop.

Northrop staff are now working 24 hours a day, but cannot handle enough parts simultaneously to stay on schedule. The increased number of workers is adding to the project's overall cost, the US Government Accountability Office reported last month.

There are other problems, too. In April 2017, a technician applied too high a voltage during a test, damaging components of the propulsion system that took more than a month to replace. In October, engineers discovered several tears in the sunshield caused by “workmanship error”. And part of the five-layer sunshield snagged during a deployment test.

In a statement, Northrop said that it “remains steadfast in its commitment to NASA and ensuring successful integration, launch and deployment” of the telescope.

All the testing is crucial because the JWST will operate from an orbit about 1.5 million kilometres from Earth, where it cannot be serviced by astronauts as the Hubble Space

Telescope was. The JWST will be 100 times more powerful than Hubble, and will survey the Universe mainly in infrared wavelengths. Among the many celestial phenomena that it aims to explore are the first stars and galaxies to form in the Universe, as well as planets in and beyond the Solar System.

NASA had asked scientists to submit proposals for the JWST's first set of observations by next week, but the agency has cancelled that deadline. “We'd rather have it launch later and work perfectly than rush and have problems,” says Emily Levesque, an astronomer at the University of Washington in Seattle. “But there are going to be a lot of people considering what this means for astronomy.”

Garth Illingworth, an astronomer at the University of California, Santa Cruz, says that the next decadal survey, scheduled for 2020, should be postponed. “It's hard to imagine a group of people thinking clearly about what to do in the future with uncertainty about JWST's performance hanging around,” he says.

NASA will also have to figure out how to accommodate the extra costs — perhaps by taking them out of the operations budget for the JWST, penalizing Northrop or delaying WFIRST. The agency has spent \$7.3 billion on the JWST so far, Lightfoot said, and cannot exceed \$8 billion without permission from Congress. ■

BIOMEDICAL RESEARCH

Cancer researchers push to relax rules for clinical trials

US government examines whether study criteria unnecessarily exclude some people.

BY HEIDI LEDFORD

Nearly 20% of publicly funded cancer clinical trials in the United States fail because investigators are unable to enrol enough participants. Yet patients and their physicians often grow frustrated when they encounter the sometimes-insurmountable requirements to join a study.

Now, researchers are pruning the lengthy lists of eligibility criteria for trials, in the hope of nixing unnecessary rules that might be hindering research. On 16 April, representatives of the US Food and Drug Administration (FDA) will meet stakeholders in Washington DC to discuss how restrictive eligibility criteria for clinical trials could be limiting people's opportunities to access experimental treatments — and the quality of the data generated by these

studies. The agency plans to use the information it gathers to develop guidelines for drug makers.

“You can have the greatest ideas and the greatest science,” says Stuart Lichtman, an oncologist at Memorial Sloan Kettering Cancer Center in New York City. “But if no one goes on the study, what good is it?”

Eligibility requirements are typically intended to protect either the participant or the study. Participants with some degree of liver failure, for example, might not be allowed to take part in a trial of a drug thought to pose a risk to that organ. Criteria might also exclude people with conditions that could confound the results of a study.

But some researchers say that a ‘cut-and-paste’ mentality has increased clinical-trial requirements over time, as scientists have

used previous trial protocols as templates for their next studies. That might be needlessly restricting participation in trials.

David Gerber, a lung-cancer specialist at the University of Texas Southwestern Medical Center in Dallas, and his collaborators have found that 80% of clinical trials sponsored by the US National Cancer Institute excluded people with previous cancer diagnoses (D. E. Gerber *et al.* *J. Natl Cancer Inst.* **106**, dju302; 2014). Yet in many cases, he says, the previous cancer might have been caught early and removed successfully before the person developed lung cancer.

“What really frustrates me are instances when, in my mind and in my heart, it really seemed that the patients should be eligible,” says Gerber. “If I had the exact same treatment outside of a clinical trial, I would give



Participants in clinical trials of cancer drugs must often meet a lengthy list of eligibility criteria.

it to them without a concern.”

A joint project by the FDA, the American Society of Clinical Oncology (ASCO) in Alexandria, Virginia, and the advocacy group Friends of Cancer Research in Washington DC has found that five common criteria for cancer-trial eligibility could often be amended without harming participants or the integrity of the trial. The team published its results last October (E. S. Kim *et al.* *J. Clin. Oncol.* **35**, 3737–3744; 2017).

People with HIV, for example, were once excluded from trials because of their poor prognosis. Now, with treatment, they often live

as long as people without the virus and should be included in many cancer trials, the group concluded.

The team also recommended that in some cases, researchers should ease restrictions on people with organ dysfunction. That could be particularly important in light of the ageing populations in some countries, including the United States, says Lichtman. The restrictions were put in place when cancer treatments were more broadly toxic, he notes, and might not be necessary for the more targeted drugs available today.

One recommendation that could generate

some controversy, he says, is a push to lower the age of eligibility for many adult cancer trials from 18 to 12. This reflects an understanding of basic drug metabolism, says Edward Kim, an oncologist at Atrium Health in Charlotte, North Carolina, who chaired the ASCO effort. “There is nothing magical about 18,” he says. “Your body pharmacologically metabolizes drugs the same way at age 12 as it does at age 18.”

But some adult-cancer physicians might feel uncomfortable treating younger people, and often treatment of these individuals takes place in specialized children’s hospitals, unlike adult clinical trials. Furthermore, most adolescent cancers are rare, and they can differ from adult cancers — even when they start in the same organ. This means the change might have little impact on research overall, says paediatric oncologist Peter Adamson of the Children’s Hospital of Philadelphia in Pennsylvania. But it could still help individual adolescents who might otherwise have been excluded from trials, he adds: “It’s the right thing to do.”

Kim and others are now working to see their changes implemented, and have submitted their suggestions to an influential programme that coordinates clinical development of new therapies at the US National Cancer Institute. Kim says he has been contacted by researchers at large pharmaceutical companies who are eager to make the changes in their upcoming trials.

The result, he says, could be data that are more relevant to the people whom he and his colleagues treat every day. “These patients have these characteristics and they’re going to be treated eventually by their doctors,” says Kim. “This is the real world.” ■

ASTROPHYSICS

Dark-matter detector in Italy strikes again

Upgraded experiment sees a beguiling data fluctuation.

BY DAVIDE CASTELVECCHI

A group of physicists says that it is still detecting signs of dark matter — the mystery substance thought to make up 85% of matter in the Universe — 20 years after it saw the first hints of such a signal.

DAMA, a collaboration of Italian and Chinese researchers, has announced long-awaited results from six years of data-taking, which followed an upgrade to the experiment in 2010. The findings are a boost for

the multiple groups attempting to reproduce DAMA’s results, which have been controversial and contradict those of other experiments. But DAMA’s improved sensitivity also makes its results harder to explain, physicists say.

Observations of galaxies and of the Universe’s primordial radiation imply that the vast majority of matter is of a type that is invisible and interacts almost exclusively through gravity. Many theories exist for explaining the nature of this dark matter, and lots of experiments have been attempting to detect it

through its subtle interactions with ordinary matter.

Rita Bernabei, a physicist at the University of Rome Tor Vergata who has led DAMA since its early days, presented the latest results on 26 March at a meeting at central Italy’s Gran Sasso National Laboratory, where the experiment sits in a cavern under a mountain. Like many detectors, DAMA aims to measure the tiny amount of energy given off when atoms of ordinary matter on Earth interact with unseen particles in a ‘halo’ of dark matter thought to envelop the Milky Way.

DAMA works by recording flashes of light that occur inside crystals of sodium iodide when subatomic particles hit the nucleus of a sodium or iodine ion. Interactions with dark-matter particles should make that signal vary throughout the year. That’s because, as the Sun moves around the Galaxy, Earth ploughs through the dark-matter halo more quickly in some parts of its orbit around the Sun than in others. The signals should peak in early June and be at their lowest in early December, says Katherine ▶

► Freese, a theoretical astroparticle physicist at the University of Michigan in Ann Arbor, who was part of the team that first proposed looking for such a signal, in 1986 (A. K. Drukier *et al. Phys. Rev. D* **33**, 3495–3508; 1986).

When DAMA first announced that it had seen such a fluctuation in 1997, soon after an early version of the experiment was turned on, the physics community was sceptical. Critics doubted that this effect was a genuine sign of dark matter. Instead, they said, terrestrial sources or quirks in the apparatus might be mimicking a real signal. There was also a possibility that the blip would vanish after parts of the detector were replaced with newer technology. But that didn't happen. "The modulation is still there, loud and clear," says Freese.

A number of increasingly sophisticated experiments that should also see dark matter — although using different techniques — have so far found none. But the DAMA team has continued to see a fluctuation. The group confirmed that it had seen the signal in 2013 (R. Bernabei *et al. Eur. Phys. J. C* **73**, 2648; 2013), with a previous incarnation of the experiment. The latest findings from DAMA come as other experiments attempt for the first time to corroborate or disprove the claim using the same type of sodium iodide crystal as in DAMA.

Leading that pack is COSINE-100, a US and South Korean experiment at the Yangyang underground laboratory in South Korea. Hyunsu Lee, a physicist at the Institute for Basic Science in Daejeon, says that had DAMA's signal disappeared in the new data, it would have dampened motivation for carrying out further sodium iodide experiments.

"For us, these results are very encouraging," says Susana Cebrian, a physicist at the University of Zaragoza in Spain who works on



The DAMA experiment in Italy is hunting for signs of dark matter.

another replication attempt, called ANAIS, in the Canfranc Underground Laboratory in the Pyrenees.

UNEXPECTED DEVIATION

But DAMA's latest results have a twist. The upgrade has made the detector sensitive to lower-energy collisions — signals from slower-moving particles. For typical dark-matter models, the timing of the fluctuations, as seen from Earth, should reverse below certain energies: "It should peak in December and be at a minimum in June," says Freese. The latest results don't show that.

The deviation "is refreshing, and food for thought," says Juan Collar, an experimental physicist at the University of Chicago in Illinois who works on dark-matter detection.

But many physicists still express scepticism.

Dan Hooper, a physicist at the Fermi National Accelerator Laboratory in Batavia, Illinois, tweeted on 26 March: "I cannot come up with a viable model that can produce this signal."

Freese, who isn't part of the DAMA collaboration, is more sanguine. She says that the data at low energies are still tentative, and could yet be compatible with a flip.

"It is more urgent than ever that an independent experiment based on the same technique, like ANAIS, could reproduce the effect," Cebrian says. Other experiments are planned in Australia and Japan.

Although DAMA's latest upgrades removed some potential concerns that the effect might have been generated inside the detector, Collar says: "The mystery, however, remains of why their result is incompatible with just about every other finding in this field." ■

S. SCHIAVON/LNGS-INFN

POLICY

Copyright reforms draw fire from scientists

Planned changes to EU regulations prompt concerns that they will impede open science.

BY QUIRIN SCHIERMEIER

A n influential committee of the European Parliament is due to vote this month on changes to copyright regulations, but the latest drafts of the rules have triggered a wave of criticism from open-science advocates. They say that the proposals will stifle research and scholarly communication.

Intellectual-property experts agree that

existing EU copyright rules need an overhaul for the digital age, and a proposal first circulated by the European Commission in 2016 had this goal in mind. But critics worry that some provisions in more-recent proposals for the law — known as the directive on copyright in the digital single market — conflict with Europe's principles of open science and freedom of expression.

"Copyright law must not hamper open science," says Vanessa Proudman, European

director of the Scholarly Publishing and Academic Resources Coalition (SPARC), a science-advocacy group in Apeldoorn, the Netherlands. "The EU has made significant headway towards open access of research funded by European citizens. The proposed new rules would clearly impede further progress, threatening the visibility of Europe's research," she says.

Concerns focus on a provision that would let publishers claim royalties for the use of snippets

of information, such as tables or headlines. This was included with the aim of enabling news publishers to secure revenue from social-media platforms such as Facebook and Google. But a proposal added by a European-parliament committee would mean that the provision also applies to academic publications.

Many scholarly publishers, including the International Association for Scientific, Technical and Medical Publishers (STM), based in Oxford, UK, support this amendment. But open-research advocates say that facts and information in a scientific article must remain free from copyright. “We really don’t want further paywalls on top of any research materials libraries have paid for already,” says Maria Rehinder, a copyright specialist in Aalto, Finland, with the Association of European Research Libraries.

FEE CONTROVERSY

Some researchers express concern that the proposed rule might even force scientists to pay fees to publishers for references they include in their own publications. But STM “cannot envisage any situation where students and researchers would need to pay fees” for citations, says Matt McKay, a spokesperson for STM.

The EU copyright law, as written, would also compel research repositories to prevent uploads

of copyrighted papers and other content. Currently, the onus is on academic publishers to issue take-down notices for papers illegally posted to repositories.

The scholarly social network ResearchGate, for example, has in recent months disabled public access to more than 1.7 million papers on its site, in compliance with take-down messages by publishers. This process of removing

“We really don’t want further paywalls on top of any research materials libraries have paid for already.”

articles upon request, says Proudman, works well and effectively for institutional repositories. Forcing all existing non-profit educational and research-data services, including more than 1,000 university repositories, to seek copyright licences and install upload filters would overburden most institutions, she says. “The proposed level of surveillance would put science repositories in the same boat as Facebook or YouTube,” she says, by requiring them to scan submissions for possible copyright violations.

The proposed rules aren’t all bad news for science, says Marie Timmermann, who is in charge of EU legislation and regulatory affairs at Science Europe, an association of national research-funding agencies in Brussels.

articles upon request, says Proudman, works well and effectively for institutional repositories. Forcing all existing non-profit educational and research-data services, including more than 1,000 uni-

Text-mining — in which researchers use computer programs to extract data automatically from large numbers of texts — is exempted from the copyright law, when carried out in the public interest. Scientists at public research organizations would be allowed to harvest facts and data from all sources they have legal access to read.

However, this exemption does not extend to companies — a possible problem for EU-funded research projects, which increasingly include commercial partners, Timmermann notes.

The European Parliament legal committee’s vote on the law, scheduled for 23–24 April, will be a crucial test of whether lawmakers are listening to scientists’ concerns. The precise version the committee will consider has not yet been finalized and circulated, and the final law will also need to be approved by the entire parliament and by EU member states before it can come into effect, due for next year. “For the sake of European research, we hope the worst flaws will yet be deleted,” Timmerman says. ■

CORRECTION

The Editorial ‘AI diagnostics need attention’ (*Nature* **555**, 285; 2018) gave an inaccurate description of the methods in a 2017 study. The model detected breast cancer in whole slide images, not mammograms.

CLARIFICATION

The News story 'Copyright reforms draw fire from scientists' (*Nature* **556**, 14–15; 2018) should have made it clear that when Vanessa Proudman talked of “that process” she was referring to how institutional repositories deal with copyright violations.



Send in the germs

Lab mice are usually kept squeaky clean, but some immunologists think a dose of dirt could make them more useful for science.

BY CASSANDRA
WILLYARD

On an unseasonably warm February morning, Mark Pierson takes a 20-minute drive to one of Minneapolis's larger pet shops. Pierson, a researcher in an immunology laboratory at the University of Minnesota, often comes here to buy mice, so most of the staff know him. Today he asks for ten, and an employee fishes them out of a glass box. Pierson requests the smaller mice because they're typically younger, but he isn't too picky. They probably all have what he wants: germs.

These mice are about to enter one of the most tightly controlled labs in the country, a facility normally reserved for studying dangerous pathogens such as tuberculosis and chikungunya virus. The rodents probably don't carry serious human infections, but they do harbour

ILLUSTRATION BY GOSIA HERBA

diseases that pose a grave threat to the hundreds of other research mice in the building.

The pet-shop mice are about to get new room-mates. Each one will bunk with a group of shiny black lab mice, sharing food, water, bedding and, most importantly, pathogens. Until now, the lab mice have been kept in a squeaky clean environment, free from most diseases, so some will fall ill and die. The rest will develop more robust immune systems, more like those of wild mice — and, arguably, humans.

What Pierson is doing breaks the rules. For more than 50 years, scientists have worked to make lab mice cleaner. In most labs today, the animals' cages are sanitized, and their water bottles and food are sterilized. "We really go to great lengths to keep natural infectious experience out of the mouse house," says David Masopust, an immunologist at the University of Minnesota who heads the lab where Pierson works. Those efforts have paid off: with the confounding effects of pathogens controlled, mouse experiments have become less variable.

But a raft of studies now suggests that this cleanliness has come at a cost, leaving the rodents with stunted immune systems. In a quest for standardized and spotless mice, scientists have made the creatures a less-faithful model for human immune systems, which develop in a world teeming with microbes. And that could have serious implications for researchers working to usher treatments and vaccines out of the lab and into the clinic. Although it's not yet possible to pin specific failures on the impeccable hygiene of standard mouse models, Masopust thinks the artificial environment must have some effect. It's no secret that the success rate for moving therapies from animal to humans is abysmal — according to one estimate¹, 90% of drugs that enter clinical trials fail. "You have to wonder if you might sometimes get misinformed simply because you're in a clean environment," says Masopust.

"Is this a mouse issue, or is this really just a lab-mouse issue?"

That's why he and other researchers are developing dirtier models that better replicate how the immune system develops in the natural world. Some groups have given their mice infections^{2,3}, others a more natural microbiome^{4,5}. But housing the dirtier mice can be risky. Pet-shop mice carry so many infections, it's as if they came from "a Dickensian orphanage," says Aaron Ericsson, a microbiome researcher at the University of Missouri in Columbia. Lab-animal caretakers take biosecurity very seriously and mice are a precious resource. "The last thing you'd want to do is have some sort of an outbreak."

DISH THE DIRT

Masopust began thinking about the cleanliness problem more than a decade ago. He was struck by how much the immune make-up of lab mice differs from that of humans. At the time, many researchers blamed the differences on genetics, but Masopust suspected that lab mice are different in part because of where they live. "Is this a mouse issue," he wondered, "or is this really just a lab-mouse issue?"

To answer that, Masopust started comparing the immune systems of lab mice to those of mice he had trapped in barns and bought from pet shops. Lab mice had many fewer cancer- and infection-fighting memory T cells — immune cells that have previously been exposed to pathogens — in their blood. They were also almost entirely lacking T cells in other tissues in the body. Humans, wild mice and pet-shop mice are swarming with these tissue-resident memory T cells. Overall, the lab mice's immune systems looked less experienced, more like that of a human infant than that of an adult.

Masopust suspected that past infections played an important

part. If so, he thought he might be able to induce changes in the lab mice's immune systems by exposing them to infectious agents. If the lab-mouse problem was cleanliness, could he make them dirtier?

He devised a seemingly simple experiment: he would drop a pet-shop mouse into a cage with several lab mice. The lab mice would pick up whatever the pet-shop mouse was carrying — everything from fur mites and pinworms to mouse hepatitis — and perhaps become immunologically more like the pet-shop mouse. This co-housing approach would let the researchers "take our cherished well-defined inbred strains and push them closer to the kind of normal immune experience that a human would have," says Stephen Jameson, a University of Minnesota immunologist who collaborates with Masopust.

But there was one major hurdle: the researchers had nowhere to put the germ-ridden rodents. "The last thing I want to do is contaminate my colleagues' mouse colonies," Masopust says. When he first discussed the experiment with the animal-resources staff, "it definitely induced heart palpitations," he says. But in a stroke of good luck, the university was about to construct a high-containment laboratory in Masopust's own building. The facility was designed for biosafety-level-three (BSL-3) research, meaning that it would securely contain pathogens that can infect humans. But it would also prevent mouse pathogens from spreading to other mice. In 2013, Masopust and his colleagues managed to secure a room there. "I was lucky," he says. "It was under-utilized. They needed revenue. That helped them be open-minded." Today, that room houses 500 mice in plastic cages, each one containing a handful of sleek lab mice and one scrappy pet-shop mouse.

After a month bunking with the pet-shop mice, the newly dirty lab mice had many of the same immunological features as wild and pet-shop mice². They had more differentiated memory T cells than normal lab mice, and they developed tissue-resident memory T cells.

The standard lab mice looked immunologically similar to newborn babies in terms of which of their genes were more or less active, but the gene-activity profiles of pet-shop and co-housed mice were closer to those of adult humans. The dirty mice also mounted much greater resistance than clean mice when the researchers infected

them with the bacterium *Listeria monocytogenes*: three days after infection, the number of bacteria they were carrying fell by more than four orders of magnitude, a response comparable to that of lab mice that have been vaccinated against the bacterium.

Soon after Masopust began work in the BSL-3 lab, Herbert Virgin, an immunologist at Washington University in St. Louis, Missouri, and his colleagues independently embarked on a similar project to understand the immune systems of lab mice. But rather than using pet-shop mice to transmit infections, they decided to transmit the infections themselves, an approach that offered more control than co-housing. "As somebody who also has trained as a virologist, I like to know what the pathogen is, going in," says Tiffany Reese, a member of Virgin's lab at the time, and now a viral immunologist at the University of Texas Southwestern Medical Center in Dallas.

They selected four pathogens: two types of herpesvirus, one influenza virus and an intestinal worm called a helminth that chronically infects the small intestines of mice. The pathogens were all similar to those that often infect children in developing countries. The researchers gave the mice the infections one at a time and allowed the animals time to recover before administering the next infection — in much the same way as humans get an infection, recover, then get another. Another group of mice received mock inoculations with saline. The final immune challenge was a vaccination against yellow fever, which uses a live but weakened form of the virus.

Like Masopust's group, the researchers noticed significant changes in the sequentially infected mice³. They differed in their gene-expression profiles and in their response to the vaccination: at first, both

groups had the same antibody responses, but a month later, the co-infected mice had lower antibody levels. It's not clear yet whether this difference affected how well the vaccine worked. "I think the jury is out about whether it has any specific utility," Virgin says. Still, he hopes that these dirtier models will lead to greater mechanistic understanding of the immune system.

CALL OF THE WILD

Other researchers have bypassed the pet shop in their quest for dirty mice. Immunologist Stephan Rosshart at the US National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD) in Bethesda, Maryland, has driven hundreds of kilometres, visiting horse barns throughout that state and the District of Columbia to collect wild mice.

Rosshart had joined the lab of NIDDKD immunologist Barbara Rehermann in 2013, and the two began poring over the literature on the microbiome, the collection of microorganisms that live on and in a larger organism. The studies showed that the microbiome has a huge influence on the immune system, but most of the papers they found were based on a comparison of two types of lab mouse: some with a lab-derived microbiome and others with no microbiome at all. What would happen, Rosshart wondered, if he gave a lab mouse a wild microbiome? That would preserve the mouse's genetic background but push its physiology closer to that of its wild cousins.

Rosshart had specific requirements for his wild-microbiome donor: he wanted an adult, genetically similar to a lab mouse and free of pathogens so it didn't risk infecting other mice at the US National Institutes of Health (NIH). "I tried to convince Stephan that that's a very bad research idea because it's very difficult," says Rehermann. But Rosshart could not be dissuaded. So each morning, he drove to between 3 and 10 barns, emptied more than 100 mouse traps and drove back to NIH with the mice. He then dissected them and preserved their tissue and faeces. In the evening, he retraced his route, collecting even more mice and baiting new traps with peanut butter. His days began at 4:30 a.m. and ended around midnight. He followed this routine seven days a week for two months. "When you do this for, like, one week it's fun, but after a while it gets very challenging," he says.

By the end, Rosshart had handled more than 800 mice. He and his colleagues selected three with the right genetics and no sign of pathogens. They transferred microbes from the animals' faeces to pregnant germ-free mice. When those mice gave birth, they passed this microbiome to their pups. The team compared this group with germ-free mice that had a microbiome derived from the sanitized lab environment.

Then they infected the mice with a mouse-adapted flu virus; 92% of the wild-microbiome mice survived, compared with just 17% of mice with the standard lab microbiome⁴. The wild-microbiome mice also developed less-severe disease when the researchers exposed them to chemicals that cause colon cancer. "The provocative hypothesis is that if you make a mouse more like a real mouse in the natural world, this becomes a better model for humans who also live in the natural world," Rehermann says.

More wildness doesn't always lead to greater infection-fighting power, however. Last month, Andrea Graham, an evolutionary ecologist at Princeton University in New Jersey, and her colleagues showed that letting lab mice re-wild themselves makes them more susceptible to worm infections⁵. Graham gave her lab mice free run of eight outdoor enclosures. When she released the first batch, they immediately began exploring the enclosure, digging burrows and sampling new food. "They were blissed out. They pulled a couple of all-nighters," she says. The microbes they encountered significantly affected the mice's ability to control some types of parasite. Mice in Graham's lab tend to clear parasitic infections rather quickly. But outdoors, "within a couple of weeks they had huge worm burdens," she says. The researchers are still trying to unpack why that might be, which could help to reveal how the immune system works in a more natural environment. Perhaps the system prioritizes fighting deadly microbes — viruses and

bacteria — over less-fatal infections such as worms, says Rosshart. "The immune response cannot be perfect against everything," he adds.

The dirty models have generated a great deal of excitement. "In many ways, they are landmark studies," says Alexander Maue, head of microbiome products and services at Taconic Biosciences, a breeder and supplier of lab animals based in Rensselaer, New York. These dirty mice, he says, will allow researchers "to look at different mechanisms of protective immunity that you wouldn't find in the normal mouse model".

MODELS FOR THE MASSES

But researchers don't yet know which models will work best for which research questions. In Masopust's version, for example, each group of lab mice gets a different cocktail of pathogens. That's both a curse and a blessing, Masopust says, because humans are variable, too. In Virgin's design, the mice get a defined set of pathogens, but the impact on the immune system isn't quite so robust.

Eleanor Riley, an immunologist at the University of Edinburgh, UK, says none of these models can fully replicate what happens in nature⁶. Wild mice differ from lab mice in many ways: diet could play a part, or sex, daylight or temperature. "I think we need to work more with ecologists and zoologists and look at the real world," she says. "There is a danger of taking a slightly reductionist, simplistic approach."

Even recreating such a simplistic version of the wild in a lab is a headache, says Virgin. "I don't think people have any question that this is important, but actually doing the experiments requires a lot of infrastructure." The wild-microbiome model gets around many of the problems of working with pathogens, but as Rosshart well knows, catching wild mice comes with its own challenges.

Whether dirty mouse models represent the human condition better than standard lab mice — and provide a better testing ground for drugs — also remains to be seen. The ideal experiment would involve taking a therapy that failed in clinical trials and retesting it in the new models to see whether the results match what happened in humans.

That's exactly what Masopust's group is doing, working with two drug companies. One has a therapy that failed in human studies, and the company would like to know whether the dirty mice could have predicted that failure. Another asked Masopust to use his mice to test a candidate therapy that works well in clean mice. The preliminary data suggest that it does not have much of an effect in dirty mice.

Colonies of dirty mice are springing up in other places. Daniel Campbell, an immunologist at the Benaroya Research Institute in Seattle, Washington, received a grant from the NIH last December to set up his own collection. He and his colleagues want to test treatments they have developed for autoimmunity, in which the immune system starts attacking healthy tissues. Therapies for such conditions seem to work well in pathogen-free mice. But "a lot of those have not translated real well into humans", he says. Campbell thinks dirty mice, which have a more developed immune system than standard lab mice, might be a more realistic model in which to test those therapies. For example, they might allow researchers to better detect unwanted side effects. "The concern is safety," he says.

Campbell says that getting the co-housing model up and running has been challenging, but he thinks the results will be worth the trouble. And many of his colleagues have questions that they'd like to test on dirty mice once the colony is ready. "I think there's a lot of interest," he says. "I think they'll all want in." ■

Cassandra Willyard is a freelance science journalist based in Madison, Wisconsin.

1. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. *Nature Biotechnol.* **32**, 40–51 (2014).
2. Beura, L. K. et al. *Nature* **532**, 512–516 (2016).
3. Reese, T. A. et al. *Cell Host Microbe* **19**, 713–719 (2016).
4. Rosshart, S. P. et al. *Cell* **171**, 1015–1028.e13 (2017).
5. Leung, J. M. et al. *PLoS Biol.* **16**, e2004108 (2018).
6. Abolins, S. et al. *Nature Commun.* **8**, 14811 (2017).

COMMENT

CLIMATE Nations vulnerable to global warming should steer geoengineering research **p.22**

HEALTH Jim O'Neill hails Hans Rosling's swansong on the predictive power of data **p.25**



SATIRE Happy 90th, Tom Lehrer, songwriter and mathematician **p.27**

OBITUARY Günter Blobel, protein-postcode Nobel laureate, remembered **p.32**

ERIC LAFFORGUE/ART IN ALL OF US/GETTY



To reduce people's consumption of added sugar, various governments have introduced a tax on sweetened drinks.

Reward food companies for improving nutrition

Governments must provide incentives for businesses to fix the global food system, not just punish them for acting irresponsibly, argues **Lawrence Haddad**.

This month, the UK government introduces the Soft Drinks Industry Levy. Producers, packagers and importers of beverages that contain 5 grams of sugar or more per 100 millilitres will have to pay a tax. Some might increase their prices to cover the cost, which could discourage buyers. The hope is that most firms will make their products less sweet to avoid it.

More than 20 countries now apply some variant of a 'sugar tax'. Various studies show that it can reduce people's consumption of added sugar^{1,2}. After Mexico's government

introduced a tax on sweetened drinks in 2014, for instance, sales in 2015 fell by nearly 10% (ref. 3).

Such 'sticks', policies that punish food and drink companies for harming people's nutrition, are popular with governments, United Nations agencies, non-governmental organizations and others. But in my view, a fundamental — and often justified — distrust of industry means that those trying to fix food systems are missing opportunities to encourage private-sector businesses to do more good things for nutrition, not

just fewer bad things. We should use 'policy carrots' too.

HISTORY OF DISTRUST

Over the past few decades, many of those who work to improve people's nutrition have seen businesses as part of the problem, not as an essential part of the solution. Much of the wariness stems from how companies have promoted breast-milk substitutes and sugary drinks.

Since 1981, the International Code of Marketing of Breast-milk Substitutes ▶

► has sought to protect the exclusive breastfeeding of infants younger than six months, and to position it as a complement to other foods for older infants. Adopted by the World Health Assembly, the decision-making body of the World Health Organization (WHO), the code aims to shield mothers, health workers and health-care systems from commercial promotion that undermines breastfeeding.

Yet producers in some countries often violate this code, for example by encouraging health facilities to include formula milk in the packs given to new mothers or by offering it free or discounted to pregnant women⁴.

The marketing and lobbying techniques used by some producers of sugary drinks to target children are similarly scandalous. Examples include branding educational materials, embedding advertisements for unhealthy food in computer games, or using toys to market such foods to children in restaurants⁵. Such drinks significantly increase people's risk of developing type 2 diabetes, heart disease and other chronic conditions. Some producers also refuse to take at least some responsibility for the rise in obesity throughout Latin America, Africa and Asia — a trend that correlates with an upswing in the consumption of soft drinks in these regions over the past 15 years.

These flashpoints in nutrition probably explain why policy carrots are rarely deployed, despite numerous studies indicating their potential value⁶. Of the countries that informed the WHO about their fiscal policies to promote healthy diets in 2016–17, more than half had increased taxes on unhealthy foods and beverages. Less than one-quarter had introduced subsidies to lower the cost of healthier alternatives⁷.

But businesses are the main investors in the world's food systems. In 2016, Hershey and General Mills each spent more than US\$500 million on advertising alone (see go.nature.com/2u3jttr). In 2014, international aid donors spent just \$50 million in total on combating diet-related chronic disease⁸ (see 'Top investors').

Punitive policies, government guidelines on eating healthily and legislation for food safety won't be enough to alter food systems such that more people are better nourished. Governments must also give incentives to the main investors in such systems so that they play a much more positive part in improving nutrition.

WORKING TOGETHER

For 25 years, I worked solely in the public sector. I now direct a non-governmental organization that supports public–private approaches to promote the availability, affordability and desirability of nutritious food — the Global Alliance for Improved Nutrition (GAIN) in Geneva, Switzerland.

Just 18 months in the job has convinced me that many in the private sector are willing to adjust their businesses to make money and improve people's nutrition at the same time.

Some heads of companies have their own reasons, such as diet-related chronic disease in their family. For other companies, dedicating resources to causes such as nutrition can pull in talented and driven employees. Overall, I have been struck by the commitment, knowledge and integrity of many in this newly discovered private-sector 'tribe'.

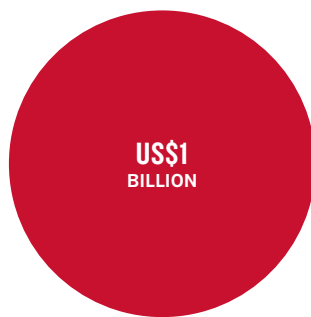
Public–private collaborations could improve nutrition in many ways. The top 10 multinationals produce more than 50% of all soft drinks. But the top 10 'packaged food' companies — which supply branded products sold in shops and supermarkets — account for only 15% of these sales in the world⁹. So those in the public sector should work with small, medium and large national companies, not just the vast multinationals. Partnerships could even involve companies that aren't in the food sector.

Mobile-phone companies, for instance, can send people government-approved text messages about how to eat healthily, or spread links to information about healthy diets. By providing this public service, they attract more customers¹⁰. Such an approach is being tried in Tanzania, Bangladesh and Ghana with the coordination of GSMA, the membership organization for more than 300 mobile-phone providers.

Likewise, marketing and advertising companies can help those in the public sector to improve the 'stickiness' of media messaging around nutrition. As an example, in 2014,

TOP INVESTORS

Globally, companies have much deeper pockets for food promotion than have governments and non-governmental organizations.



Advertising spend of just 2 of the top 25 US food and beverage companies (Hershey and General Mills) in 2016.

US\$50
MILLION



International aid invested in nutrition-related diseases in 2014.

the Indonesian government collaborated with a creative agency in Indonesia, GAIN and the London School of Hygiene and Tropical Medicine. The result was a one-minute video during which a mother gossips about how everyone else is failing to feed their children properly. The use of humour and emotion seemed to work, in contrast to standard government-produced instructions about what people should be eating. An independent evaluation of the campaign indicated that it helped 50% of the 6- to 23-month-old infants in the assessed villages to meet a nutrient adequacy threshold, compared with 36% of infants in the control villages¹¹.

Companies that specialize in food transport or packaging can help to reduce food loss during storage and distribution using relatively low-cost technologies, such as repurposed storage containers or cheap insulating materials. (Perishable foods such as fruit tend to be higher in micronutrients than longer lasting ones such as cereals.) Last year, led by the Lagos state government, we at GAIN connected tomato farmers in Nigeria with commercial suppliers of reusable plastic crates. Studies in Asia have shown that such crates, used in place of wicker baskets, can reduce the loss of fruits and vegetables along the supply chain from 30–50% to 5% (ref. 12).

With a mix of public- and private-sector technical and financial assistance, small- and medium-sized businesses in, say, horticulture and aquaculture could make their products more available, affordable, desirable and profitable. Since 2013, GAIN has been working with around 500 such firms to get more servings of nutritious foods (such as beans, fish, peanuts and chicken) into markets in five countries in Africa and Asia, and to make those servings cheaper. Independent evaluations show some achievements. For example, one firm in Kenya has helped to make tilapia fish affordable for 68% of the population (up from 49%) in the region where it is operating¹³.

FIVE STEPS TO BETTER FOOD

So how do governments and others that are striving to improve nutrition identify and seize opportunities to change behaviours? Five things need to be done.

Support businesses that work with nutritious foods. Governments frequently create export-processing zones or business parks with reduced rents or tariffs for exported goods, say, to promote business types that boost economic growth. So why not create business parks for producers of nutritious foods, with lower rents and taxes and cheaper electricity and water supplies?

Most of the small- and medium-sized businesses aiming to become the next



Swapping wicker baskets for crates provided by packaging firms better protects tomatoes in transit.

food giants find it hard to access financial services. These companies are too high-risk to attract investment from banks and tend to be ineligible for microfinance schemes that provide small loans (around \$1–10) to very poor households. Governments could create financial instruments, such as low-interest loans for nutritious-food suppliers, to meet their needs.

Governments could also develop effective ‘quality seals’ targeted to specific groups, such as street-food vendors or institutional caterers, to certify that a food is healthy (or unhealthy). Assessments suggest that ‘traffic-light systems’ of red, amber and green ratings used in the United Kingdom and Australia, or the black stop-sign labels used in Chile, seem to be effective ways to steer people towards better nutrition¹⁴.

Create demand for healthy foods. Business leaders often tell me that if consumers wanted more nutritious foods, they would meet that demand. Yet businesses shape demand, and some bend it towards unhealthy foods — mainly because these are easy to produce at scale and to transport, market and sell at a significant mark-up.

Governments must take the lead when it comes to building consumer demand for healthy foods — much as they have changed people’s behaviour around smoking and drink-driving. That means partnering with non-profit foundations and creative agencies to make health messages about food accurate and memorable instead of worthy and dull. In the United Kingdom, the Food Foundation, a non-profit organization working to improve nutrition, collaborated with the creative agency ifour in late 2017 to create messages and images that tap into children’s interest in superheroes to encourage them

to eat vegetables. The impact of this has not yet been rigorously evaluated, but a related initiative targeting households in a province of Ecuador in 2015 increased egg consumption to one a day in 6- to 9-month-olds, and improved growth¹⁵. Providing such incentives will boost consumer demand and thus encourage companies to meet it.

Create models to emulate. Both governments and businesses need evaluated examples of things they can do together that work.

Much of the evidence for the effectiveness of public–private partnerships comes from other sectors, such as health, infrastructure and education; from the unpublished reports of public and private organizations;

“Governments can make it hard for businesses to do good things for nutrition.”

or from the minds of those involved. A 2016 literature review concluded that “there are few independent, rigorous assessments

of the impact of commercial sector engagement in nutrition”.¹⁶

UN agencies, non-governmental organizations, businesses, researchers and nutrition champions need to do a better job of disseminating the lessons learned, for instance by creating a knowledge repository. This could be similar to the World Bank’s free-to-access website on public–private partnerships for the building of roads, ports and other infrastructure (see go.nature.com/2ptdgqm).

Name and fame — or shame. Businesses can derail public-health initiatives and distort publicly available research to suit their own ends. For instance, a 2015 investigation by the *British Medical Journal* found that researchers at UK advisory bodies had

received funding from major soft-drinks companies¹⁷. How exactly this affects research is not yet known, but it clearly undermines trust.

Likewise, governments can make it hard for businesses to do good things for nutrition — either through a lack of awareness or through poor planning. For example, some governments impose a tariff on imported premix, the micronutrient-rich compound that is used in small amounts to fortify staple foods, such as wheat or maize (corn). The tariff can dissuade food processors from implementing this cost-effective public-health strategy.

A ranking scheme is needed to flag which governments and businesses are doing positive or harmful things for nutrition.

One game changer has been the Access to Nutrition Index (www.accesstonutrition.org). Released every three years, the index uses mainly self-reported information to evaluate the world’s 22 largest multinational food and beverage manufacturers on their policies, practices and performance in relation to under-nutrition and obesity. Although it has begun to produce national reports, further national, independent and evidence-based assessments are needed.

The World Bank currently ranks 190 national economies according to how easy it is to start and operate a firm in that country. In principle, governments could similarly be ranked according to how easy they make it for businesses to produce available, affordable and desirable nutritious food that can, for instance, reduce the proportion of women experiencing anaemia or the percentage of children who are obese. Such assessments (perhaps conducted jointly by the World Bank, the WHO and the Food and Agriculture Organization of the UN) would help to reveal what kinds of government action actually help businesses to improve nutrition.

Foster public–private engagement. More dialogue between people working on nutrition in the public and private sectors will catalyse all these other steps.

Major differences in culture, language and networks exist between those concerned with food systems in the public and private sectors. In fact, before I joined GAIN, I talked and worked only with academics, programme implementers and policymakers.

The accountability measures I describe could help those in the public sector to decide who to partner with. Also, thorough and pragmatic conflict-of-interest guidelines will help to reveal when public-health goals are at risk.

There are numerous ways to foster more dialogue. Panels at conferences should include participants from both sectors. Public funders could incentivize joint

proposals for nutrition research from public–private collaborations. Companies and public-sector organizations could set up staff exchange programmes. And executive-level courses, either at universities or in private institutions, could bring together professionals from both sectors to learn from instructors drawn from these two worlds.

Many analysts (myself included, in the past) have drawn parallels between ‘big tobacco’ and ‘big food’. In both cases, major corporations wield immense power over consumers and society, and their products are capable of doing considerable harm.

But there are crucial differences. Unlike big tobacco, big food is not the only player. There are small- and medium-sized companies too. And big tobacco cannot make tobacco that promotes public health, whereas big food can and does produce nutritious, sustainable foods. Motivated by both carrots and sticks, the industry can produce more — at a lower price. ■

Lawrence Haddad is executive director of the Global Alliance for Improved Nutrition (GAIN) in Geneva, Switzerland.
e-mail: lhaddad@gainhealth.org

1. Briggs, A. D. M. et al. *Lancet Public Health* **2**, e15–e22 (2017).
2. Nakhimovsky, S. S. et al. *PLoS ONE* **11**, e0163358 (2016).
3. Colchero, M. A., Rivera-Dommarco, J., Popkin, B. M. & Ng, S. W. *Health Aff.* **36**, 564–571 (2017).
4. Save the Children. *Don't Push It: Why the Formula Milk Industry Must Clean Up Its Act* (Save the Children, 2018).
5. Cairns, G., Angus, K., Hastings, G. & Caraher, M. *Appetite* **62**, 209–215 (2013).
6. Afshin, A. et al. *PLoS ONE* **12**, e0172277 (2017).
7. World Health Organization. *Global Nutrition Policy Review 2016–2017 (DRAFT)*. (WHO, 2018).
8. International Food Policy Research Institute. *Global Nutrition Report 2016: From Promise to Impact: Ending Malnutrition by 2030* (IFPRI, 2016).
9. Alexander, E., Yach, D. & Mensah, G. A. *Glob. Health* **7**, 26 (2011).
10. Turner, T., Spruijt-Metz, D., Wen, C. K. F. & Hingle, M. D. *Pediatr. Obes.* **10**, 403–409 (2015).
11. University of Sydney Impact Evaluation Consortium for Global Alliance for Improved Nutrition. *Effectiveness of an Integrated Program to Reduce Maternal and Child Malnutrition in Indonesia: Cross-Sectional Impact Evaluation Report* (Global Alliance for Improved Nutrition, 2017).
12. Lipinski, B. et al. *Reducing Food Loss and Waste. Working Paper* (World Resources Institute, 2013).
13. Altai Consulting. *USAID–GAIN Case Study Annex to Technical Report on the MNF Assessment* (USAID/GAIN, 2016).
14. Cecchini, M. & Warin, L. *Obes. Rev.* **17**, 201–210 (2016).
15. Iannotti, L. L. et al. *Pediatrics* **140**, e20163459 (2017).
16. Hoddinott, J. F., Gillespie, S. & Yosef, S. *World Rev. Nutr. Diet.* **115**, 233–238 (2016).
17. Gornall, J. *Br. Med. J.* **350**, h231 (2015).



Developing countries must lead on solar geoengineering research

The nations that are most vulnerable to climate change must drive discussions of modelling, ethics and governance, argue **A. Atiq Rahman** and colleagues.



A group of villagers stands beside the Jamuna River in Bangladesh, where erosion is eating into the riverbanks.

essential. As the scale of the damage grows, more countries will turn to the “loss and damage” provisions in the Paris agreement. And these are vague: who should pay how much, and to whom, for lost farming or fishing livelihoods? What size of cheque would compensate for the destruction of coral reefs?

In that context, solar geoengineering — injecting aerosol particles into the stratosphere to reflect away a little inbound sunlight — is being discussed as a way to cool the planet, fast. The technique is controversial, and rightly so. It is too early to know what its effects would be: it could be very helpful or very harmful. Developing countries have most to gain or lose. In our view, they must maintain their climate leadership and play a central part in research and discussions around solar geoengineering.

HIGH STAKES

Solar geoengineering is outlandish and unsettling. It invokes technologies that are redolent of science fiction — jets lacing the stratosphere with sunlight-blocking particles, and fleets of ships spraying seawater into low-lying clouds to make them whiter and brighter to reflect sunlight. Yet,

“There is a limit to what populations threatened by sea-level rise, biodiversity loss, droughts and hurricanes can do.”

if such approaches could be realized technically and politically, they could slow, stop or even reverse the rise in global temperatures within one or two years. No other way of doing this has been conceived.

Removing greenhouse gases from the air would take decades, if it is even possible.

A decade of modelling research indicates that solar geoengineering might reduce many of the worst effects of climate change if deployed in moderation. For example, injecting 5 megatonnes of sulfur dioxide into the stratosphere — about one-quarter of that released by Mount Pinatubo’s eruption in 1991 — each year could keep warming below 2 °C. (However, there are likely to be limits to how much cooling can be achieved, especially under high greenhouse-gas emissions scenarios⁵.) Studies have found that solar geoengineering should also be able to reduce climate impacts on hydrology, redressing trends in which wet regions get wetter and dry regions get drier⁶. Lower temperatures would slow global sea-level rise⁷ and could curb the increasing incidence and strength of tropical cyclones⁸.

A decade ago, there were serious concerns that solar geoengineering might produce stark winners and losers and might disrupt the monsoons. Research has allayed these worries. For example, it seems conceivable that moderate solar geoengineering would

People in the global south are on the front line of climate change. As global temperatures creep upwards, the Intergovernmental Panel on Climate Change (IPCC) is forecasting rising seas eroding small island states¹, declining food production in many regions of Asia², water stress across Africa³ and major loss of biodiversity in South America⁴.

Developing countries have spoken out on climate policy. Links between climate justice and development are now accepted, as is the idea that nations have common responsibilities — emitters are liable for impacts

felt elsewhere. Despite having emitted very little greenhouse gas themselves, the world’s least-developed countries and small-island states demanded that the 2015 Paris climate agreement require warming to be kept “well below” 2 °C, and that a 1.5 °C limit should also be explored.

But there is a limit to what populations threatened by sea-level rise, biodiversity loss, droughts and hurricanes can do. Mitigation of climate change is crucial. The emissions cuts agreed in Paris are not enough — they will take the world to a 3 °C rise (see go.nature.com/2u3ybkh). Adaptation is therefore

benefit many regions that are vulnerable to climate change, with few losers. Monsoon rains would be affected less than if climate change proceeds unchecked⁹.

But solar geoengineering is no panacea; it could compound some risks of climate change. It would only mask the warming effect of greenhouse gases. Ocean acidification would still pose a threat to marine life if carbon-dioxide emissions were not slashed. Sulfur dioxide might delay ozone regeneration in the stratosphere. And whichever aerosol was used to filter out sunlight, more research would be needed on its impacts on health and the environment.

The overall effects of solar geoengineering are uncertain. All studies so far are based on computer simulations, which are poor at forecasting regional climates, for example. The Earth system might hold surprises that digital models do not capture. The projections require thorough and sceptical examination.

Furthermore, solar geoengineering raises difficult socio-political issues that cannot be wished away. It is uncertain how, or whether, the technique could be governed in ways that ensure prudence, accountability and justice. Who has the right to implement an inherently global technology? Would the technology weaken multilateral commitments to reduce emissions such as the Paris agreement?

These issues matter deeply to developing nations. But most solar-geoengineering research is being done in the well-heeled universities of Europe and North America. Unless that changes, voices from the global north will set the policy agenda and decide which research projects should be accelerated or shut down.

We are neutral on whether solar geoengineering should ever be used. It has not yet been established whether it would be a beneficial addition to meeting the Paris goals. We recognize its potential physical risks and socio-political implications. And we oppose its deployment until research into its safety and effectiveness has been completed and international-governance mechanisms established. But we are committed to the co-production of research and to well-informed debate.

Others have already taken sides. Some people in the global north have tried to convince their peers in the south that they should reject solar geoengineering. Campaigners who vehemently oppose it often make their case by emphasizing the risks and playing down the potential benefits¹⁰. We take issue with this paternalism and propose an inclusive way forward.

BIG DECISIONS

Developing countries must be in a position to make up their own minds. Local scientists, in collaboration with others, need to conduct research that is sensitive

to regional concerns and conditions. For example, what effects might solar geoengineering have on hurricanes in the Caribbean, flooding in Bangladesh or agriculture in East Africa? Broader discussions among academics, policymakers, the public and public intellectuals are needed on climate risks and justice.

To begin this process, we (and the co-signatories of this Comment) have been running solar-geoengineering engagement workshops across the global south — the first of their kind — as part of the SRM Governance Initiative (SRMGI), in which SRM stands for solar radiation management. International and non-governmental, SRMGI was launched in 2010 by the Royal

“Solar geoengineering raises difficult socio-political issues that cannot be wished away.”

Society in London, The World Academy of Sciences (TWAS) in Trieste, Italy, and the Environmental Defense Fund in New York City. The regional workshops — held mostly in the past

three years in Bangladesh, Brazil, China, Ethiopia, India, Jamaica, Kenya, Thailand, New Zealand (for the Pacific states), Pakistan and the Philippines — have brought together local climate scientists, journalists, policymakers and representatives of civil society to learn about and discuss solar geoengineering.

Participants had no consensus position on the technology. But they raised common hopes and concerns. In general, we found widespread opposition to deployment at this stage, but support for studies of local impacts. As a participant at the Nairobi workshop put it: “This idea is crazy ... but we have to understand it.” Many were sceptical about whether the methods would work and if developing countries, rather than more powerful governments, would have any say in how and whether solar geoengineering is deployed.

To fund regional research, this week, SRMGI issues the first call for applications to a US\$400,000 fund called Developing Country Impacts Modelling Analysis for SRM (DECIMALS). The fund is administered by TWAS and financed by the Open Philanthropy Project, a foundation backed by Cari Tuna and Dustin Moskovitz (co-founder of Facebook and the project-management app Asana). Developing-world scientists can apply to DECIMALS for funds to model the solar-geoengineering impacts that matter most to their regions. International collaborations will be supported and researchers will be asked to run local workshops to promote wider discussion of the implications of their findings.

Further outreach and research in the developing world will require extra support from governments, universities and civil society

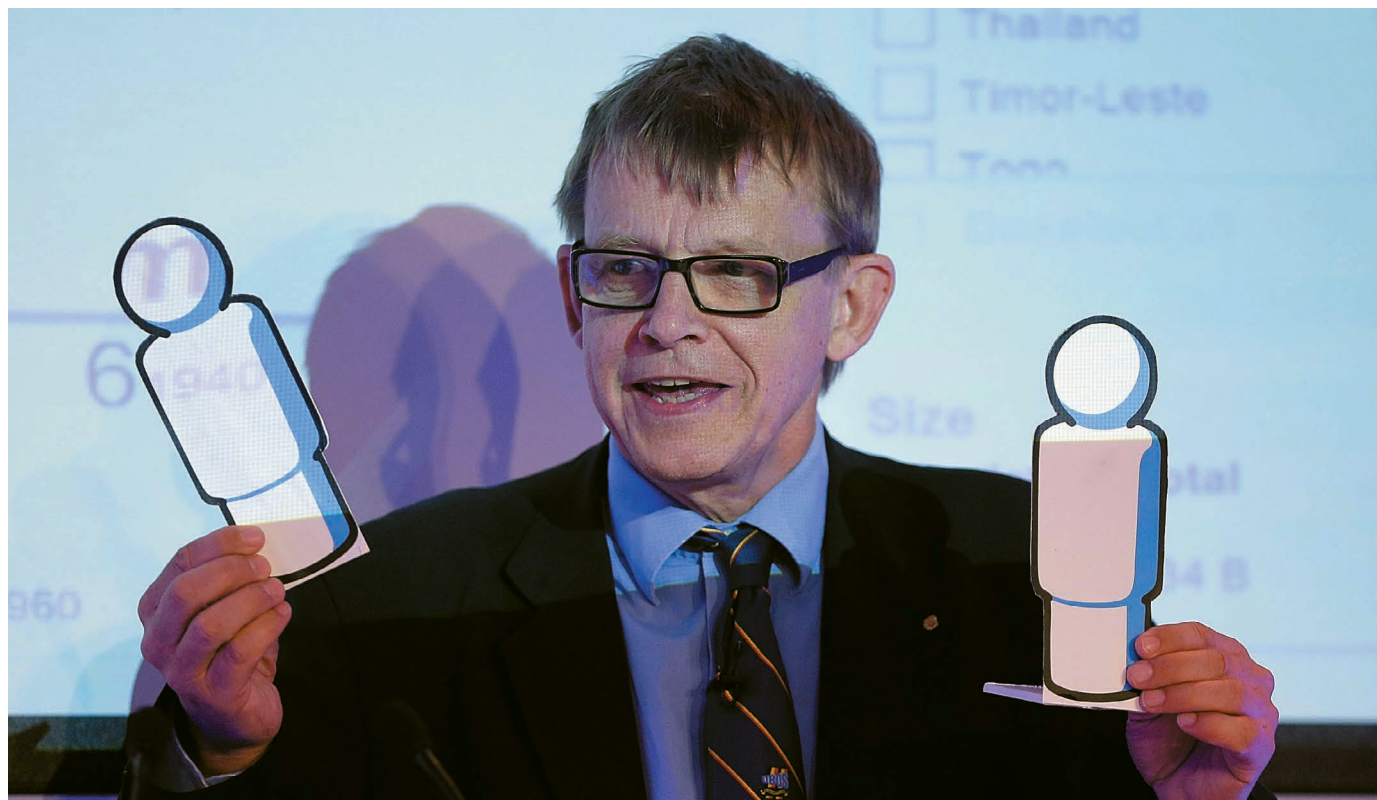
worldwide. Research funders in advanced economies should fund collaborations with scientists in developing countries. We would like to see an IPCC special report on the risks and benefits of solar geoengineering. Ultimately, a coordinated global research initiative — perhaps under an organization such as the World Climate Research Programme — is needed to promote collaborative science on this controversial issue.

Solar geoengineering is fraught with risks and can never be an alternative to mitigation. But it's unclear whether the risks of solar geoengineering are greater than the risks of breaking the 1.5°C warming target. As things stand, politicians will face this dismal dilemma within a couple of decades. It is right, politically and morally, for the global south to have a central role in solar-geoengineering research, discussion and evaluation. ■

A. Atiq Rahman is executive director of the Bangladesh Centre for Advanced Studies, Dhaka, Bangladesh. **Paulo Artaxo** is professor of environmental physics, Institute of Physics, University of São Paulo, São Paulo, Brazil. **Asfawossen Asrat** is professor of geology, School of Earth Sciences, Addis Ababa University, Addis Ababa, Ethiopia. **Andy Parker** is project director of the Solar Radiation Management Governance Initiative and honorary senior research fellow, School of Earth Sciences, University of Bristol, Bristol, UK. The authors write on behalf of 8 co-signatories. e-mail: aparker@srmgi.org
A.P. declares competing financial interests.

1. Nurse, L. A. et al. Small islands. In *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Part B: Regional Aspects (eds Barros, V. R. et al.) 1613–1654 (Cambridge Univ. Press, 2014).
2. Hijioka, Y. et al. Asia. In *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Part B: Regional Aspects (eds Barros, V. R. et al.) 1327–1370 (Cambridge Univ. Press, 2014).
3. Niang, I. et al. Africa. In *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Part B: Regional Aspects (eds Barros, V. R. et al.) 1199–1265 (Cambridge Univ. Press, 2014).
4. Magrin, G. O. et al. Central and South America. In *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Part B: Regional Aspects (eds Barros, V. R. et al.) 1499–1566 (Cambridge Univ. Press, 2014).
5. Kleinschmitt, C., Boucher, O. & Platt, U. *Atmos. Chem. Phys.* **18**, 2769–2786 (2018).
6. Curry, C. L. et al. *J. Geophys. Res. Atmos.* **119**, 3900–3923 (2014).
7. Moore, J. C., Jevrejeva, S. & Grinsted, A. *Proc. Natl Acad. Sci. USA* **107**, 15699–15703 (2010).
8. Moore, J. C. et al. *Proc. Natl Acad. Sci. USA* **112**, 13794–13799 (2015).
9. Reynolds, J., Parker, A. & Irvine, P. J. *Earth's Future* **4**, 562–568 (2016).
10. ETC Group. *Geoengineering and Climate Change: Implications for Africa* (ETC Group, 2014).

A full list of co-signatories and details of competing financial interests accompany this article online (see go.nature.com/2pjbevu).



Hans Rosling discusses population growth at the ReSource 2012 conference in Oxford, UK.

STATISTICS

Swansong of a data visionary

Jim O'Neill hails the last book by Hans Rosling, the statistician who recast progress.

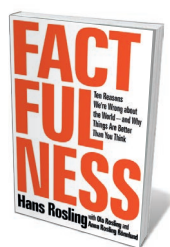
Hans Rosling was many things. A physician and epidemiologist; a statistician and data visualizer; a staunch advocate of free data as the bedrock of an accurate world view. *Factfulness* is Rosling's last, and posthumous, book; he died in February 2017. Like his other work, including his famous presentations, it throws down a gauntlet to doom-and-gloomers in global health by challenging preconceptions and misconceptions.

Co-written with Ola Rosling and Anna Rosling Rönnlund (Rosling's son and daughter-in-law), *Factfulness* is a fabulous read, succinct and lively. It asks why so many people — including Nobel laureates and medical researchers — get the numbers so wrong on pressing issues such as poverty, pandemics and climate change. The book isolates the ten instincts that lead to what Rosling calls the “overdramatic worldview”: that pervasive, generally pessimistic global perspective that often cancels out significant progress made in the face of vast challenges.

The “gap instinct”, for instance, is the tendency to divide everything into

two — such as developed and developing countries. This ‘us and them’ construct, he shows, is not borne out by the facts: some 75% of people live in middle-income countries. Rosling adds nuance to the global socio-economic picture by outlining four income levels, from US\$1 a day to more than \$64. He offers a masterclass in avoiding the gap instinct by, for example, comparing averages — one number can hide a range.

Many of the other instincts that Rosling examines reveal his acuity about how easy it is for us to slide into negative, fearful or grandiose mindsets when faced with monolithic problems. His discussion of the “blame instinct”, for example, is important because it



Factfulness: Ten Reasons We're Wrong About the World — and Why Things Are Better Than You Think
HANS ROSLING WITH OLA ROSLING AND ANNA ROSLING RÖNNLUND
Flatiron: 2018.

reminds us of people's impatience with complex causation and tendency to pin everything on one agent. As he notes, problems such as lack of research into the diseases of the poorest people are systematic. To change them, you have to understand the systems.

Rosling enlivens the chapters with multiple-choice questions. These highlight our knee-jerk responses and the scale of progress in a number of areas. (He jokes that chimpanzees choosing at random would generally outperform humans on these.) For example, he asks: if there are 2 billion children in the world today, how many will there be by 2100 — 4 billion, 3 billion or 2 billion? I got this wrong. Another question is how many of the world's one-year-old children have been vaccinated against at least some form of disease: 20%, 50% or 80%? This I got right, but few people do. These riveting exercises kept me up late one evening: I couldn't wait to see how stupid I was with the next question.

Rosling also covers five risks that we “should worry about”: global pandemic, financial collapse, a new world war, climate change and the extreme poverty that still ▶

► afflicts 700 million people. *Factfulness* is not, however, comprehensive. I have wondered whether Rosling would have been his usual cheery self when confronted with the challenges of antimicrobial resistance (AMR), and the needed solutions. I had the chance to discuss this issue with him only briefly. Its absence from the book is disappointing, but then, I'm biased. I led a global review on AMR for the UK government under then-prime minister David Cameron, during which I departed from my own generally cheery take on the state of the world.

I came to Rosling's work in the past decade, when someone at Goldman Sachs Asset Management (which I chaired) suggested that I should follow one of Rosling's online presentations about the state of the world, its evolution and probable future. It was wise advice. Although I initially

sulked because of the inference that Rosling discussed such information better than me, I soon realized that he took aspects of what I had become

"I loved Rosling's positive approach to life's challenges; it did, and does, inspire me."

known for to a completely new level. His knowledge of issues around health, disease control and many aspects of development was obviously deep and broad. A few years later, I shared a platform with him. I loved his positive approach to life's challenges; it did, and does, inspire me.

Long before US President Donald Trump and fake news, I warned audiences about slavishly believing what they read. As Rosling spells out, our awareness of facts has never been that strong, and many journalists are not motivated to tell us that the world is getting better. (No news is sometimes the only good news.) It is up to individuals to ensure that we foster the disciplined habits of mind that Rosling eloquently and clearly sets out.

This magnificent book ends with a plea for a factual world view. Rosling was optimistic that this outlook will spread, because it is a useful navigational tool in a complex world, and a genuine antidote to negativity and hopelessness. A just tribute to this book and the man would be a global day of celebration for facts about our world. Perhaps Trump should lead the charge on that. ■

Jim O'Neill is an economist, a cross-bencher in the UK House of Lords, a distinguished visiting fellow at the think tank Chatham House in London, and honorary chair of economics at the University of Manchester, UK.

NEUROSCIENCE

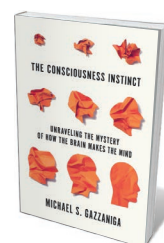
From meat to mind

Douwe Draaisma enjoys Michael Gazzaniga's exploration of the biological basis of consciousness.

In April 1648, a young admirer of René Descartes visited the French philosopher at his estate on the Dutch coast. Aiming to discuss the cardinal points of Cartesian philosophy, Frans Burman had marked more than 70 passages in Descartes's works. How, he asked, can soul and body affect each other, given their fundamental difference? Descartes conceded that the question was thorny, but pointed to the evidence that they do — for instance, in emotion. The mystery lies in the mechanism, and this, Descartes confided, was perhaps best left to theologians.

Neuroscientist Michael Gazzaniga tackles this abiding mind-body problem anew in *The Consciousness Instinct*. His subtitle, *Unraveling the Mystery of How the Brain Makes the Mind*, rephrases Descartes's conundrum into a bold promise. But then, Gazzaniga is a bold scientist. He made his name in the 1960s through pioneering work on severing the connection between the brain's left and right hemispheres ('split brains'), as his autobiography vividly details (D. Draaisma *Nature* 518, 298–299; 2015).

His latest book is certainly evidence that scholars advancing in age (Gazzaniga is now 78) often trust themselves with ever broader scientific and philosophical questions. Thus he guides readers through neurology, biology and psychology, discussing the origin and neural underpinnings of language or the mechanism of facial recognition.



The Consciousness Instinct: Unraveling the Mystery of How the Brain Makes the Mind

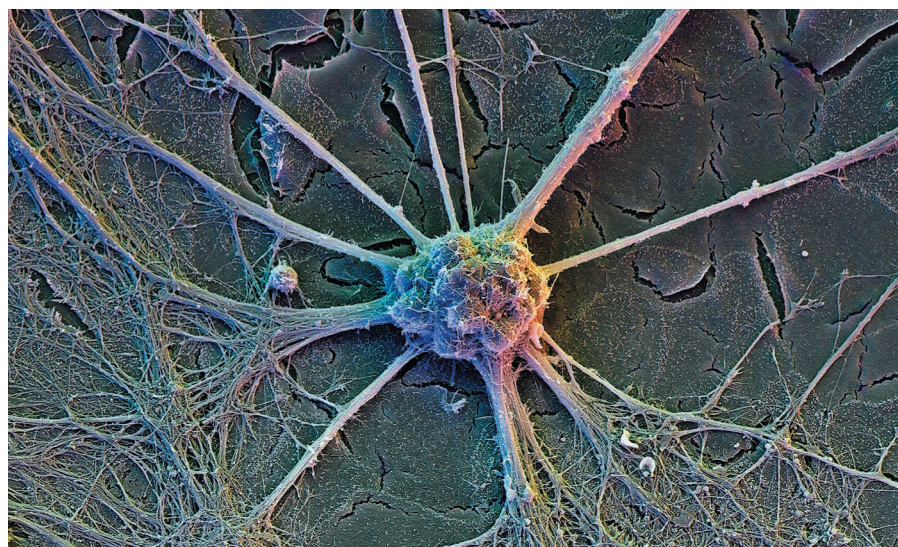
MICHAEL S. GAZZANIGA
Farrar, Straus and Giroux: 2018.

And he evokes Isaac Newton's laws of motion, the special and general theories of relativity and quantum physics — surprisingly, for a man with a self-confessed blind spot for mathematical abstraction.

The tour yields a couple of useful lessons. With theoretical biologist Howard Pattee, Gazzaniga emphasizes that we should resist the lure of the "single-explanation fallacy" — the idea that one theory can cover

everything, from our introspective sense of awareness down to the subatomic particles of brain tissue. Explanations, he asserts, should be thought of as context-dependent, just as light in quantum physics sometimes behaves like waves and sometimes like particles.

Gazzaniga defines consciousness as "the subjective feeling of a number of instincts and/or memories playing out in time in an organism". He points out that clinical cases — he spent a few years working on neurological wards — add complexities. For instance, people who are completely unable to move can still be conscious, a frightening condition



A nerve cell in the human brain, seen in a false colour under a scanning electron microscope.

DAVID SCHARF/SPL

called locked-in syndrome. Consciousness might be absent in sleepwalking. Thus, coupling it to behaviour is misleading.

Nor is it straightforward to link consciousness to parts of the brain. One of Gazzaniga's earliest findings was that disconnecting the left and right hemispheres produced two separate conscious systems; only one, usually supported by the left brain, was able to express itself in language. It had been assumed that consciousness co-evolved with the cerebral cortex, supporting 'higher' functions such as language and reasoning. But referring to the work of neuroscientist Björn Merker, Gazzaniga makes the case that consciousness might not be necessarily — or exclusively — locked into cortical and linguistic processes. In some children born with a seriously compromised forebrain, the damaged tissue gets replaced by fluid (hydranencephaly). They grow up lacking language, but still express feelings and have subjective experiences. According to Gazzaniga, consciousness might actually originate in the evolutionarily older midbrain, with the cortex providing "a collection of extensions (apps!) to enhance conscious experiences".

In an engaging discussion of the brain's architecture, he offers a mundane simile for consciousness. The brain should be thought of as a multitude of modules, each specialized for a single task, such as recognizing patterns or monitoring rhythm in music. The end products of these modules rise to the surface and burst like "bubbles in a boiling pot of water", each a fleeting part of our awareness. Our subjective sense of continuity, described by pioneering psychologist William James as "stream of consciousness", might be illusion: we merely experience the rapid succession of elements as a smooth movement, like the frames of a film. The metaphor of the bubbles seems first and foremost an invitation to generate a testable theory, and Gazzaniga's observations will almost certainly provide much of the test material.

Gazzaniga ends by reflecting that the ultimate explanation for how mind emerges from meat might not prove "warm and cuddly". Instead, it might vie with quantum mechanics for sheer counter-intuitive weirdness, hovering "way beyond our intuitions and imaginations". Once again we seem to hear what Burman heard, 370 years ago: a sigh of resignation, as Descartes indicated that it might all be better left to the theologians. ■

Douwe Draaisma is professor of the history of psychology at the University of Groningen in the Netherlands.
e-mail: d.draaisma@rug.nl



Tom Lehrer performing in San Francisco, California, in 1965.

MATHEMATICS

Tom Lehrer at 90

Andrew Robinson looks back at the scientific high notes in the mathematician and satirist's inimitable oeuvre.

In 1959, the mathematician and satirist Tom Lehrer — who turns 90 this month — performed what he characteristically called a "completely pointless" scientific song at Harvard University in Cambridge, Massachusetts. (He was a PhD student there at the time.) 'The Elements', now one of his most cherished works, sets the names of all the chemical

elements then known to the tune of the 'Major-General's Song' from *The Pirates of Penzance*, the comic opera by W. S. Gilbert and Arthur Sullivan. Lehrer's heroically precise, rapid-fire enunciation of 102 elements (reordered to allow flawless end-rhymes), ends with the much-quoted crack, "These are the only ones of which the news has come to Harvard/And ▶

► there may be many others but they haven't been *discovered*."

In the 1960s, Lehrer followed up with more than a dozen astrigent, cynical and often pointedly political songs, such as 'So Long, Mom, I'm Off to Drop the Bomb (A Song for World War III)'. As *The New York Times* had it, "Mr. Lehrer's muse [is] not fettered by such inhibiting factors as taste." (Lehrer reprinted the quote in his album liner notes.) In the fraught geopolitics and paranoia of the cold war, however, Lehrer's social criticism touched a chord with many in the United States. Fans might, however, have been surprised to learn that he had crunched numbers for the National Security Agency as an army draftee in the mid-1950s.

Much of Lehrer's oeuvre — some 50 songs (or 37, by his own ruthless reckoning) composed over nearly three decades — played with tensions at the nexus of science and society. His biggest hit, *That Was The Year That Was*, covered a gamut of them. This 1965 album gathered together songs Lehrer had written for *That Was The Week That Was*, the US satirical television show spawned by the BBC original. 'Who's Next?' exposes the dangers of nuclear proliferation. 'Pollution' highlights environmental crises building at the time, such as undrinkable water and unbreathable air.

The rousing ballad 'Wernher von Braun' undermines the former Nazi — who designed the V-2 ballistic missile in the Second World War and later became a key engineer in the US Apollo space programme. In Lehrer's view, it was acceptable for NASA to hire von Braun, but making him into an American hero was grotesque. "Once the rockets are up, who cares where they come down? 'That's not my department,' says Wernher von Braun" — lines that still resonate in today's big-tech ethical jungle. 'New Math', meanwhile, skewers the education system through the lens of a misfired revolution in mathematics, with its telling refrain: "It's so simple, so very simple, that only a child can do it" (A. Bellos *Nature* 516, 34–35; 2014).

LYRICAL PRECISION

Lehrer — who grew up on New York City's Upper East Side — certainly sees a connection between his mathematical training, which began at Harvard at the prodigiously young age of 14, and his compositions. He was drawn to songwriting

in his teens; after failing to respond to classical-music training, he switched to the study of popular music. In an interview in 2000, he summed up the fields' dual impact. "The logical mind, the precision, is the same that's involved in math as in lyrics," he said. "It's like a puzzle, to write a song."



The cover of Tom Lehrer's debut album, released in 1953.

Lehrer agrees with mathematician Stanislaw Ulam (one of the builders of the atomic bomb) that rhyming "forces novel associations ... and becomes a sort of automatic mechanism of originality". As he told me in 2008: "If 'von Braun' didn't happen to rhyme with 'down' (and a few other words),

"The logical mind, the precision, is the same in math as in lyrics."

the most quoted couplet in the song would not exist, and in all probability the song itself would not have been written."

His musical career began at university, with the spoof sports song 'Fight Fiercely, Harvard'. In the early 1950s, Lehrer put on a satirical show in the physics department, *The Physical Revue* (a pun on the name of the US journal then named *Physical Review*). With co-performers including Norman Ramsey (later a Nobel laureate in physics) and Lewis Branscomb (who would become a presidential science-policy advisor), he performed ditties such as 'Relativity', 'Fugue for Scientists' and 'The Slide Rule Song'. It was a training ground for later triumphs.

He began recording in 1953. Although US radio stations refused to play such 'controversial' material, his fame spread through word of mouth. In Britain, the royal approval of unexpected fan Princess

Margaret and the support of the BBC significantly raised Lehrer's profile, and he considered abandoning academia. But in 1960, bored by touring, he returned to Harvard, aiming to complete a long-standing mathematics PhD on modes in statistics. Soon, however, he concluded he had nothing original to contribute academically. As he notoriously wrote in 'Lobachevsky', a song named after a nineteenth-century Russian mathematician: "Plagiarize!/Let no one else's work evade your eyes!/... So don't shade your eyes,/but plagiarize, plagiarize, plagiarize/— only be sure always to call it, please, research." Lehrer dropped his doctorate and began to teach mathematics — at the Massachusetts Institute of Technology in Cambridge in 1962 and, from 1972 until his retirement in 2001, at the University of California, Santa Cruz (along with a class in musical theatre).

He also largely gave up songwriting and public performing in the early 1970s. Following the award of the Nobel Peace Prize to then-US Secretary of State Henry Kissinger in 1973, Lehrer commented: "Political satire became obsolete." And

in 2002 he remarked, still less optimistically: "Things I once thought were funny are scary now. I often feel like a resident of Pompeii who has been asked for some humorous comments on lava." About the political earthquakes triggered by US President Donald Trump, Lehrer has been silent.

As for his songs, their vigour, concision, melodic variety and humour never stale. Although Lehrer is absurdly omitted from the *Encyclopaedia Britannica* (unlike his friend, the lyricist and composer Stephen Sondheim), his scathing creations remain one of the most original — not to mention mathematically elegant — bodies of artistic work to come out of the United States in the twentieth century. ■

Andrew Robinson is the author of *Einstein: A Hundred Years of Relativity* and *The Last Man Who Knew Everything, a biography of Thomas Young*.
e-mail: andrew@andrew-robinson.org

CORRECTION

The exhibition review 'Fake Views' (*Nature* 555, 442; 2018) erroneously described the artist Ryuta Nakajima as a woman; he is a man.

Correspondence

Hawking, Sulston and science in Europe

As well as championing open and accessible research, physicist Stephen Hawking and biologist John Sulston were part of a long tradition of British engagement with European science (for obituaries, see *Nature* 555, 444; 2018 and *Nature* 555, 588; 2018). As chief scientific advisers to the European Commission, we feel strongly that this tradition must not end — irrespective of where the future takes the United Kingdom.

Science for the greater public good depends on openness of mind, of spirit and of borders. Sulston and Hawking did much to uphold these ideals and to promote the importance of basing policy decisions on strong scientific evidence. Both recognized that society benefits from integrated scientific endeavour. Indeed, European projects built on this premise — such as the intergovernmental research organizations CERN (Europe's particle-physics laboratory near Geneva, Switzerland) and the European Molecular Biology Laboratory — have strengthened science by stimulating the movement of ideas across the continent and beyond.

These great scientists shared a strong sense of social decency and encouraged a profound respect for expertise, each of which are more important now than ever.

Rolf Heuer, Paul Nurse
European Commission, Brussels, Belgium.
rolf.heuer@cern.ch

Nobel principles hold true after 123 years

Nils Hansson and colleagues suggest that Nobel committees in 1901–66 were persuaded to award the Nobel Prize in Physiology or Medicine based on the potential research impact of a single discovery or innovation, rather than on a distinguished

research record (*Nature* 555, 311; 2018). As secretary general of the Royal Swedish Academy of Sciences, I can confirm that this is still the case.

The Nobel prize is not a lifetime achievement award. In his last will of 1895, Alfred Nobel stipulated that the Physiology or Medicine prize should go to “the person who shall have made the most important discovery within the domain”; in physics should be for “the most important discovery or invention within the field”; and in chemistry should be awarded for “the most important chemical discovery or improvement”.

It is reassuring that the assessment of hundreds of nominations by Nils Hansson (no relation of mine) and colleagues confirms that past committees have rigorously upheld Nobel's will.

Göran K. Hansson *The Royal Swedish Academy of Sciences, Stockholm, Sweden.*
goran.hansson@kva.se

How philosophy was squeezed out of PhD

Gundula Bosch's argument for putting the philosophy back into the PhD is a breath of fresh air (*Nature* 554, 277; 2018). It is interesting to look back and see how broad critical thinking came to be eased out of the doctorate, squeezing academic enquiry into narrow disciplines.

The process started in the early 1970s in the United States, prompted by a suspicion that intellectual artefacts of the ‘soft’ sciences, as they were then called — such as sociology, anthropology and philosophy — were stimulating campus unrest.

This conveniently dovetailed with the idea that if industry outsourced its research and development departments to universities by setting (and funding) curricula, then students would have ready-made jobs in industry on graduation. These mechanistic conceits looked good on paper and fitted well

with reductionists' educational metrics. However, they all but killed students' curiosity for serendipitous scientific enquiry.

My father designed stellar-inertial guidance systems for reconnaissance aircraft and, after he retired, would often present his work to physics and engineering students. When they asked him what they should study to prepare for such a career, he would reply: “Read the classics,” by which he meant Aristotle, Ralph Waldo Emerson, Jean-Jacques Rousseau and Blaise Pascal.

The best scientific and technical progress does not come out of a box. It is more likely to emerge from trying to fit wild, woolly and tangential ideas into useful societal and economic contexts.

Michael Stocker *Ocean Conservation Research, Lagunitas, California, USA.*
mstocker@ocr.org

Sciences unite for Spain's prosperity

As Spain's economy recovers, the strategic application of science could help to stimulate prosperity and to attract much-needed investment. In an unusual move in a world of specialization, the Spanish scientific community has formed a meritocratic, all-sciences advisory council within the Gadea Foundation for Science in Madrid, a non-profit body of leading scientists that works to improve Spain's science system. The council's aim is to galvanize politicians and the public into promoting research that will ensure social progress (see www.gadeaciencia.org).

Spain ranks ninth in the world for scientific production and has 58 scientists in the 2017 Clarivate Analytics Highly Cited Researchers list (see go.nature.com/2j77ctb). The application of research results for the benefit of society is still disturbingly low, however, owing to meagre public support and too few industries based on science and technology.

The advisory council's first forum was held in October 2017 to develop a strategy for improving this situation. It was framed around four cornerstones: health, life sciences (including philosophy, mathematics and astrobiology), Earth (including materials and water, food and energy, and climate change and biodiversity) and society (including science policy and the economy). The forum's founding declaration emphasizes the importance for advancing society of knowledge, training, talent and academic–industrial interaction in all of these areas. Our view is that science is not just for scientists — it is a human right.

Fernando Baquero, Jose A. Gutiérrez-Fuentes *Gadea Foundation for Science, Madrid, Spain.*
ja.gutierrezfuentes@gadeaciencia.org

Encouraging trend in US astronomy

Aswin Sekhar remarks on the low proportion of female astronomers in many countries (*Nature* 555, 165; 2018). A career in science can often exceed 50 years, meaning that the total average number of women (and minorities) will remain low for another half a century, even if we achieve parity now among early-career scientists. What matters as much as where we've been is where we're going.

As the chair of the International Astronomical Union's US committee for membership applications, I can report that we are doing much better than the grand averages would suggest. Women comprise around 40% of the latest US intake of 212 individuals, with 43% of those having gained their PhDs after 2010. This is an encouraging trend.

David Soderblom *Space Telescope Science Institute, Baltimore, Maryland, USA.*
drs@stsci.edu

Günter Blobel

(1936–2018)

Biologist who decoded how proteins are sorted in cells.

For much of the twentieth century, biologists puzzled over how the proteins that build, run and leave cells get where they need to be. Over five decades, biologist Günter Blobel hammered out the answer: the ‘signal hypothesis’, a targeting system resembling a set of postal codes. It earned him the Nobel Prize in Physiology or Medicine in 1999.

The steps to and beyond this discovery helped to explain normal cell organization, as well as many diseases in which the transport of proteins is defective. The signal hypothesis also laid the foundations of modern biotechnology. Proteins such as human insulin and human growth hormone could be tagged for mass production in bacteria, secreted and easily collected.

Blobel’s pathway to the prize began at the Rockefeller University in New York City, where he spent 50 years. As a postdoc in George Palade’s laboratory during the late 1960s, he studied the secretory pathway taken by proteins set to be extruded from the cell. He and cell biologist David Sabatini found that newly minted proteins could not be digested as they emerged from the ribosomes that synthesize them. The pair concluded that the proteins entered the secretory pathway as they were made.

Blobel and Sabatini proposed that these proteins have a distinctive region — their postal code — that is recognized by special chaperones that guide the part-synthesized protein and its ribosome to a cellular organelle called the endoplasmic reticulum. There, protein synthesis is continued as the protein crosses into the organelle and enters the secretory pathway.

Many advances in our understanding of metabolism come from breaking cells into their components, and then reassembling the parts to work out which component does what. Blobel used this approach to test his hypothesis. He mixed ribosomes, the messenger RNA that encodes proteins, cytosol and membranes *in vitro*. His experiments failed repeatedly.

But he never gave up. He continued mixing components from different organisms, convinced that, like the alchemists, if he had just the right ratios or sources of ingredients, he could create something new. Finally, in 1975, after four years in the cold room, he got a clear result: his nascent protein seemed to target and enter the secretory pathway.

As with most research, there was not one



‘aha’ moment. Blobel fretted about ugly facts that could destroy his beautiful hypothesis. Perhaps the protein did not enter the secretory pathway, it was just stuck on the outside? Maybe the tests he was using to determine whether the protein entered the pathway were faulty? Two papers, published with cell biologist Bernhard Dobberstein in the *Journal of Cell Biology* in 1975, set out crucial experimental evidence amid a flurry of controls.

The assays Blobel had developed using mainly mammalian components were quickly used to identify the signals that dispatch proteins to other cell compartments — such as mitochondria, and to chloroplasts in plant cells — and to identify similar systems in yeast and bacteria. The assays were applied to proteins to be secreted and those that would end up spanning membranes. The system was highly conserved: a mixture of human mRNA encoding a secretory protein, ribosomes from rabbit blood and membranes from yeast could synthesize the human protein. The impact of Blobel’s work cannot be overstated.

Blobel’s career was filled with audacious speculations. “I always imagined how things would be,” he once said. For him, the thrill was the experimental chase. Yet he cautioned trainees to be ready to abandon their fantasies, “when data come which aren’t compatible”. His speculations concerned the regulation of gene expression, the channels through which proteins crossed membranes and the pores that are the gateway into the cell’s nucleus.

When necessary for his work, he learnt yeast genetics, modern structural biology and electrophysiology. Inured to the ridicule of his speculations, he once said, “There is an internal revolt in me against conforming. After the war, my family lived in East Germany and that taught me that truth is the most holy and important thing in life.”

Blobel grew up in the Silesian village of Waltersdorf, where his father was a veterinary surgeon. As the Second World War was drawing to its end, when Blobel was 9, his family fled west, passing through Dresden. It was his first exposure to a big city. As an adult, he often recalled being entranced by the many spires and the magnificent cupola of the Frauenkirche. A few days later, he saw fire-bombing bring the church and city to ruin.

Blobel studied medicine in Germany, but grew frustrated that it treated symptoms rather than the causes of disease. He emigrated to the United States to do a PhD with Van R. Potter at the University of Wisconsin–Madison. There he succeeded in isolating the nucleus, which he investigated throughout his career with infectious enthusiasm. In his last years in the lab, he proposed large changes in the structures of the pores that regulate the passage of material in and out of the nucleus. He dubbed these changes ‘the ring cycle’, in tribute to the Wagnerian operas he enjoyed on rare outings from the Rockefeller campus.

Blobel was passionate about architecture — of cities as well as of cells. He donated his US\$960,000 Nobel winnings to two extraordinary projects: the rebuilding of the bombed Frauenkirche, and of the Dresden synagogue, destroyed on Kristallnacht.

Debonair and jovial, Blobel appreciated repartee and a well-told tale. His wife, Laura Maioglio, was a frequent participant in Blobel lab functions and scientific affairs at Rockefeller. And Blobel held court in her famed restaurant on Broadway, Barbetta, debating literature, design, music and, of course, science. There, with his shock of white hair and emphatic gestures, he did recall the alchemists of old. Fittingly, his wizardry transformed our understanding of the cell. ■

Sanford Simon is professor of cellular biophysics at the Rockefeller University in New York City, New York, USA, where for 34 years he tried, unsuccessfully, to disprove Blobel’s speculations.
e-mail: simon@mail.rockefeller.edu

JAMES ESTRIN/NYT/REDUX/EYEVINE



ISS CREW EARTH OBSERVATIONS/JOHNSON SPACE CENTER

Figure 1 | Holding back the flood. In 2011, the muddy floodwaters of the Mississippi were prevented from inundating Arkansas agricultural land (left) by flood defences; green areas on the right are forested regions on the floodplain. Munoz *et al.*¹ present a palaeoflood record for the Mississippi River that spans the past 500 years, and conclude that levee building and channelization have made destructive floods larger and more common than before.

HYDROLOGY

Mississippi rising

The Mississippi River is shackled by one of the world's largest systems of flood control. A palaeohydrological record suggests that those measures might actually be making floods worse. [SEE LETTER P.95](#)

SCOTT ST. GEORGE

The Mississippi River was once thought to be uncontrollable: no feat of engineering could prevent the river from bursting its banks and sending floodwaters across its natural domain. But since the late nineteenth century, an expanding system of levees, floodways and channel modifications (Fig. 1) has gradually brought the river to heel, largely confining its waters to the main channel and accelerating them downstream. On page 95, Munoz *et al.*¹ conclude that those same control measures have inadvertently raised the threat of flooding in the lower Mississippi to a level that is unprecedented in the past five centuries.

Some engineers and hydrologists have contended^{2,3} that modifications to the Mississippi, and particularly the construction of the extensive levee system, have raised the height of the water and increased the volumetric flow rate (discharge) during major floods. The earliest available report of flooding on the Mississippi⁴ described an inundation in May 1543, but the first permanent stream gauge was not established on the river until 1897, at Vicksburg, Mississippi. The discharge record from this gauge, and from several others in the river's catchment area, now includes more than a century's worth of observations. Nevertheless, because the most destructive of floods — such as the great floods that occurred in 1927, 1993 and 2011 — are relatively rare,

the perspective offered by direct hydrological measurements is still too narrow to determine, with certainty, whether the hazard posed by the river is changing.

Munoz and colleagues have worked around that limitation by building their own extended record of flooding on the Mississippi, using evidence of high waters preserved by oxbow lakes and trees. An oxbow lake forms when a meander of a river is cut off to form a free-standing body of water. When such lakes are inundated by floods, they act as natural sediment traps for sand and silt carried in floodwaters. As those particles settle, they create a layer of coarse sediment on the bottom of the lake that is distinct from the clay and fine silt left behind when the lake is not hydrologically

connected to the main channel⁵. Prolonged inundation with floodwater can cause some species of tree — particularly oak — to form wood that has abnormal features. The annual growth rings of such trees in the 'bottomland' hardwood forest of the Mississippi floodplain therefore contain a natural record of past floods⁶. By splicing together sedimentary sequences from lakes in Louisiana and Mississippi with 'flood-ring' signatures from living and dead trees in southeastern Missouri, the authors assembled a flood chronology for the lower Mississippi that stretches back to the early sixteenth century.

Together, these natural archives have kept a remarkably faithful account of past floods. The tree rings mark the occurrence of the great floods of 1844 and 1927, as well as *l'Année des Grandes Eaux* (the Year of the Great Waters) in 1785, which destroyed French–Canadian settlements in Illinois and Missouri. It is more difficult to date individual floods in the lake records because of the lower chronological resolution of that archive, but layers of coarse sediment can be matched to the great floods of 1851, 1927 and 2011, as well as to the flood reported by Spanish conquistadors in 1543. Overall, this new palaeoflood record suggests that, although flood hazards have waxed and waned through time, the Mississippi has risen higher and flooded more frequently in the past century than during any other period in the past 500 years.

The authors propose a provocative explanation for this recent hydrological intensification. Since the start of the twentieth century, records gathered using instruments show that the discharge of the Mississippi has slowly risen and fallen in concert with the surface temperature of the North Atlantic Ocean, which has alternated every two or three decades between warm and cold states⁷. Munoz and colleagues draw on proxy temperature estimates for the North Atlantic⁸ to demonstrate that this dependency has held steady since the 1500s. Because the spate of major floods in the past century cannot be explained by the observed temperature behaviour of the North Atlantic, the authors conclude that the trend towards larger and more frequent floods is mostly due to the transformation by humans of the Mississippi River and its basin.

Blaming floods on the infrastructure that was built to guard against them will be controversial. The Mississippi basin has undergone upward trends in precipitation and evapotranspiration — the sum of evaporation and plant transpiration from Earth's surface — in the past several decades⁹, so climatic factors other than the influence of the North Atlantic might also have affected the rhythm of the river. And, like the rest of North America, the Mississippi basin has warmed substantially since the end of the Little Ice Age¹⁰ (a period of cooling that began in the sixteenth century and ended in the mid-nineteenth century¹¹), so

I think it is possible that the long-term trends in hydrology could be the result of climate change, rather than river engineering. Testing these competing explanations will require more palaeoflood work to be performed along the upper Mississippi and its main tributaries, where modifications to the river are less intensive and the climate still dominates the river's hydrology¹².

In the meantime, Munoz and co-workers' study makes it clear that a century or so of hydrological readings is not sufficient to take the measure of a river such as the Mississippi. Their palaeoflood record is compelling because it offers an opportunity to step back and consider the ebb and flow of the river on a timescale that befits its majesty. And if the authors are correct, and collective efforts to subdue the Mississippi have inadvertently pushed it to rise higher than ever, then the time might have come to consider loosening its restraints. ■

Scott St. George is in the Department of Geography, Environment and Society,

University of Minnesota, Minneapolis, Minnesota 55455, USA.
e-mail: stgeorge@umn.edu

1. Munoz, S. E. *et al.* *Nature* **556**, 95–98 (2018).
2. Pinter, N., Jemberie, A. A., Remo, J. W. F., Heine, R. A. & Ickes, B. S. *Geophys. Res. Lett.* **35**, L23404 (2008).
3. Watson, C. C., Biedenbarn, D. S. & Thorne, C. R. *J. Hydraul. Eng.* **139**, 1071–1078 (2013).
4. Barry, J. M. *Rising Tide: The Great Mississippi Flood of 1927 and How it Changed America* (Simon & Schuster, 1998).
5. Toonen, W. H. J., Winkels, T. G., Cohen, K. M., Prins, M. A. & Middelkoop, H. *Catena* **130**, 69–81 (2015).
6. Therrell, M. D. & Bialecki, M. B. *J. Hydrol.* **529**, 490–498 (2015).
7. Enfield, D. B., Mestas-Núñez, A. M. & Trimble, P. J. *Geophys. Res. Lett.* **28**, 2077–2080 (2001).
8. Gray, S. T., Graumlich, L. J., Betancourt, J. L. & Pederson, G. T. *Geophys. Res. Lett.* **31**, L12205 (2004).
9. Milly, P. C. D. & Dunne, K. A. *Geophys. Res. Lett.* **28**, 1219–1222 (2001).
10. Trouet, V. *et al.* *Environ. Res. Lett.* **8**, 024008 (2013).
11. Matthews, J. A. & Briffa, K. R. *Geogr. Ann.* **87**, 17–36 (2006).
12. Wise, E. K., Woodhouse, C. A., McCabe, G. J., Pederson, G. T. & St-Jacques, J.-M. *J. Hydrometeorol.* **19**, 161–182 (2018).

ECOLOGY

Forests in flux as climate varies

How do changes in climate affect forest ecosystems? A study of temperate forests in the United States has assessed alterations in biomass and tree-species composition across a 20-year period of climate variability. SEE LETTER P.99

SEBASTIAAN LUYSSAERT
& J. HANS C. CORNELISSEN

Documenting and understanding changes induced by climate in the composition and function of vegetation is essential for planning adaptation strategies, because chances to intervene do not arise often for forests that contain trees with long lifespans. Moreover, most of the effects exerted by climate change on the composition and biomass of vegetation probably occur incrementally rather than abruptly¹, which makes their detection a challenge. On page 99, Zhang *et al.*² report an analysis of changes in forest biomass in the eastern United States over two decades, during a time when some regions became drier and others became wetter.

Ecological studies have long focused on analyses in which the main groupings for organisms being studied are determined by evolutionary relationships such as belonging to the same genus. However, such kinship can hide functional differences. For example,

even among closely related species of oak tree (*Quercus* spp.), some will thrive under moist conditions, whereas others are more suited to dry climates. In the past decade, the use of functional rather than kinship-driven approaches to grouping has provided many important insights³, and Zhang and colleagues' work can be added to the list of studies that have successfully done this.

Zhang *et al.* sought to investigate whether shifts in climate affect forest characteristics such as the prevalence of drought-tolerant species and the total biomass of the tree population. The authors compared temperate-forest inventory data⁴ gathered during the 1980s and the 2000s. This inventory includes surveys of around 100,000 plots, in which data such as species name and a standardized measurement of tree diameter were recorded for roughly 3 million trees. Diameter measurements allow the amount of biomass that is present above the surface of the ground to be estimated for a particular tree on the basis of previous studies of growth patterns for

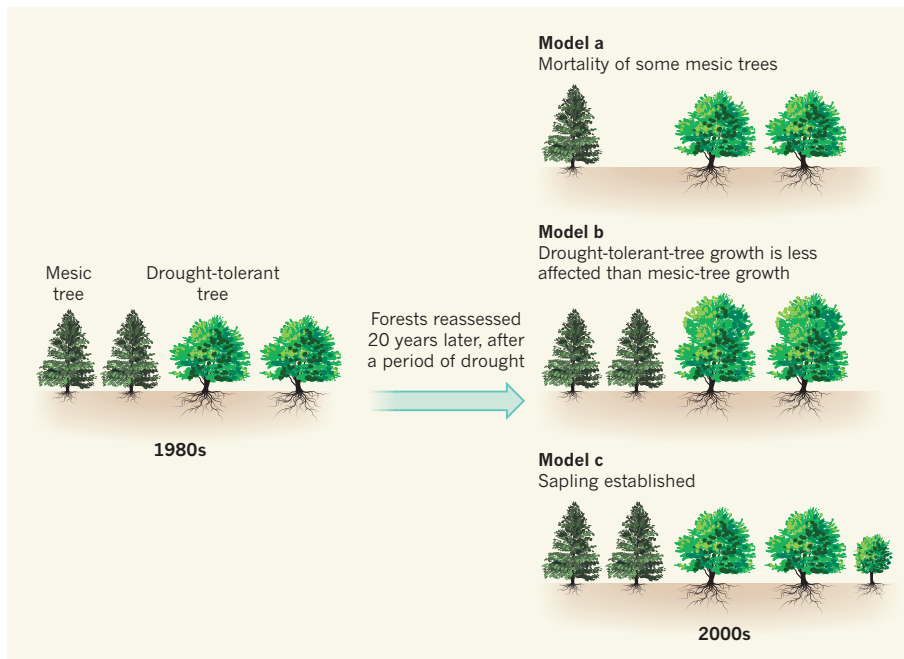


Figure 1 | Changes in the drought tolerance of tree populations. Zhang *et al.*² used an inventory⁴ of trees growing in the eastern United States in the 1980s and the 2000s to estimate how an intervening period of drought had affected the contribution of drought-tolerant species to total biomass, in forests that had reached a given age. This analysis revealed an increase in the contribution of drought-tolerant trees compared with that of moisture-needing (mesic) trees to the population-level biomass. Three models could account for such an increase. Model a is consistent with the results for forests composed of trees more than 80 years old. Model b can account for the patterns observed by the authors in most other forests. The study's time span is probably too short to fully capture the sapling emergence of model c. The drought tolerance of trees might partly depend on their entering into symbiotic interactions with root-colonizing fungi that extend the effective root length⁹. Such interactions could reduce tree biomass. If so, this might need to be considered in studies of this type.

a given species. Such analyses enabled the authors to assess the aboveground biomass per hectare for tree populations, and also to investigate whether changes occurred in the relative contribution of particular types of species, such as drought-tolerant trees, to the total aboveground biomass.

The authors analysed the data grouped into grids of cells that each covered an area of one degree of latitude by one degree of longitude (about 110 kilometres by 85 kilometres). The plots sampled in the 1980s and the 2000s were not identical; however, the authors were able to check their findings using a subset of plots that had been resampled and found that their conclusions remained the same. The authors determined the tree-age composition of the forests and assigned them into 20-year-interval age brackets. This enabled comparisons to be made between similar types of forest, for example, comparing 20–40-year-old forests in the 1980s and in the 2000s, therefore avoiding possible confounding factors such as changes in drought sensitivity that are linked to the increase in tree height as trees age⁵.

The authors assigned a numerical score to each recorded species of tree that represented its drought-tolerant characteristics. This score was generated on the basis of information about water availability, derived from

measurements of the annual precipitation at grid cells where the species was found, and the minimum water requirements of that species. For each plot, the authors calculated the average drought tolerance of its tree population by using the species' drought scores, also taking into account the relative abundance of each species. The authors then compared how drought tolerance at the tree-population level for a particular age class in a given grid cell changed in the roughly 20 years between inventories, and assessed whether these changes mirrored any changes in soil moisture (estimated by the Palmer drought severity index) at each location.

Despite the occurrence of unavoidable problems such as logging during the data-collection period, Zhang and colleagues build a strong case that, between the 1980s and the 2000s, the effects of ongoing climate variation in this zone of modest regional climate change have resulted in detectable changes in forest composition and biomass. In areas in which soil moisture increased over time, the authors observed a population-level decrease in drought tolerance and a population-level increase in aboveground biomass per hectare.

The most striking finding made by Zhang and colleagues was that, in areas in which soil moisture had decreased, a decrease in the

average aboveground biomass gained per tree was accompanied by an increase in population-level drought tolerance. This increased share of drought-tolerant species occurred because the decrease in the growth rate of moisture-needing (mesic) species during a drought is greater than the decrease in growth rate of drought-tolerant species. Consequently, the forests that reached the particular age range in the 2000s have a lower biomass and a larger proportion of drought-tolerant species than the forests that reached this age range in the 1980s.

In forests containing trees more than 80 years old that experienced drought, the population-level changes in drought tolerance observed by the authors were often driven by an increase in the mortality of mesic species (model a in Fig. 1). The role of preferential tree mortality in driving forest changes that are linked to climate variability has already been reported in a study in Europe⁶. Yet, for most age ranges, the authors found that the tree population became more tolerant to drought because the growth of the drought-tolerant species was less affected than was that of the mesic species. This means that the drought-tolerant trees increased their proportional contribution to the aboveground biomass per hectare in the study's two-decade period (model b in Fig. 1).

Although the establishment of saplings probably contributes to the response of forests to climate variability (model c in Fig. 1), the short time span of the study, the fact that forest inventories commonly measure only individual trees above a threshold size that it can take a decade or more for a sapling to reach, and the relatively small contribution of saplings to the total aboveground biomass for a given hectare of established forest, make it unlikely that Zhang and colleagues' approach would capture fully the changes in species composition that are due to sapling emergence.

The authors suggest that their results and observations could have relevance for how climate is affecting other temperate forests or forests in other climate zones. Yet perhaps the priority should be to unravel the mechanisms that underlie the connection between drought tolerance and biomass production at the species level. Zhang and colleagues suggest that the low availability of water should favour tree species that allocate a greater proportion of their biomass to fine roots, thereby promoting drought tolerance at the expense of aboveground biomass production. However, this might be only part of the climate-response phenomenon.

We speculate that, compared with mesic species, drought-tolerant species might have greater investments in symbiotic relationships with soil-dwelling mycorrhizal fungi that can colonize tree roots. Such interactions can extend the effective total root length, thereby extending access to water and nutrients

in the soil⁷. However, such connections would probably come at a substantial cost in terms of tree-biomass reduction because of the need to divert sugars to fungal partners. The type and abundance of mycorrhizal symbioses vary with soil type and climate^{8,9}, so if fungal symbiosis is a major consideration in these scenarios, such factors would need to be considered in future implementations of the approach used by Zhang and colleagues.

There is a pressing need to understand the relationship between water availability and the drought tolerance and biomass of forests. It is necessary, therefore, to ask whether the types of change that the authors observed would be able to keep pace with climate changes that

occur on longer timescales. For example, will the drought-tolerance capacity of today's saplings suffice for the conditions that these trees might encounter when they reach maturity? It's high time to knock on wood that it will, as well as to continue to investigate the mechanisms that affect forest ecosystems in a changing climate. ■

Sebastiaan Luyssaert and J. Hans C. Cornelissen are in the Department of Ecological Science, Faculty of Science, VU Amsterdam, 1081 HV Amsterdam, the Netherlands.
e-mails: s.luyssaert@vu.nl;
j.h.c.cornelissen@vu.nl

1. Parmesan, C. & Yohe, G. *Nature* **421**, 37–42 (2003).
2. Zhang, T., Niinemets, Ü., Sheffield, J. & Lichstein, J. W. *Nature* **556**, 99–102 (2018).
3. Kunstler, G. *et al.* *Nature* **529**, 204–207 (2016).
4. Bechtold, W. A. & Patterson, P. L. *The Enhanced Forest Inventory and Analysis Program — National Sampling Design and Estimation Procedures*. Gen. Tech. Rep. SRS-80 (US Department of Agriculture Forest Service, 2005).
5. McDowell, N. G. & Allen, C. D. *Nature Clim. Change* **5**, 669–672 (2015).
6. Ruiz-Benito, P. *et al.* *Glob. Change Biol.* **23**, 4162–4176 (2017).
7. Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* (Elsevier, 1997).
8. Phillips, R. P., Brzostek, E. & Midgley, M. G. *New Phytol.* **199**, 41–51 (2013).
9. Soudzilovskaia, N. A. *et al.* *Glob. Ecol. Biogeogr.* **24**, 371–382 (2015).

This article was published online on 21 March 2018.

CONDENSED-MATTER PHYSICS

Novel electronic states seen in graphene

A simple system made from two sheets of graphene has been converted from an insulator to a superconductor. The finding holds promise for opening up studies of an unconventional form of superconductivity. SEE ARTICLE P.43 & LETTER P.80

EUGENE J. MELE

In two papers in *Nature*, Cao *et al.*^{1,2} report the discovery of new electronic ground states in twisted bilayer graphene — a pair of single-atom-thick sheets of carbon atoms, stacked with their honeycomb lattices rotated out of alignment. The authors interpret one of these states² (page 80) as a correlated Mott insulator, a non-conducting state produced by strong repulsive interactions between electrons. The other¹ (page 43) is a superconductor, a state of zero electrical resistance produced by effective attractive interactions between electrons. The insulator turns into the superconductor when a small number of charge carriers are added to the graphene. This connection between the states is unlikely to be a coincidence — as Sherlock Holmes might have commented, “the universe is rarely so lazy”.

Cao *et al.* show that the stacking of graphene sheets allows access to a new family of materials with electronic behaviours that are exquisitely sensitive to the atomic alignment between the layers, which affects interlayer electron motion. This finding might surprise physicists, because electronic behaviour is usually dominated by whichever of the associated processes has the largest energy scale. But, in this case, there's a conundrum: the energy associated with electron motion between atoms within a layer is of the order of electronvolts, whereas the energy for electron

motion between layers³ is, at most, hundreds of millielectronvolts.

The resolution to this conundrum is a matter of symmetry. Well-prepared, single layers of graphene are highly ordered systems whose electronic properties are determined by a subtle symmetry, which is encoded in a solid-state version of the Dirac equation describing low-energy excitations. These excitations are

sensitive to interlayer couplings that alter the symmetries of the stack.

Interactions between electrons in these excitations can produce forms of matter generically described as being strongly correlated. A well-reasoned strategy for discovering such forms of matter has been to restrict intralayer electron motion by applying a strong magnetic field⁴. This generates narrow electron energy bands (Landau levels) in which electron–electron repulsion can control the physics of the graphene bilayer.

Cao *et al.* have taken a simpler tack to discover strongly correlated states. They used the rotational misalignment of graphene sheets to tune twisted bilayer graphene into a regime in which interactions between electrons can dominate the electronic states of the system. Such rotational misalignment forces the electronic band structures in the two sheets out of alignment and enlarges the bilayer's unit cell (the smallest repeating unit of the crystal lattice) (Fig. 1a). For large rotations, the first

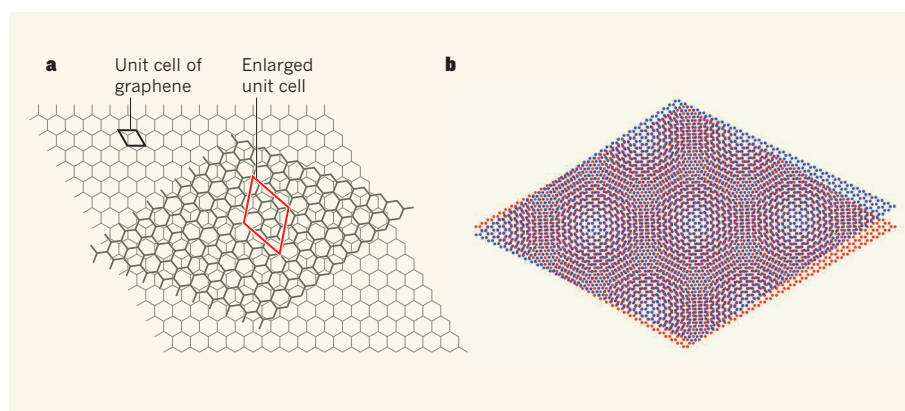


Figure 1 | The effects of rotation in twisted bilayer graphene. **a**, When a graphene bilayer is twisted so that the top sheet is rotated out of alignment with the lower sheet, the unit cell (the smallest repeating unit of the material's 2D lattice) becomes enlarged. For large rotations, the electronic band structures of the two graphene sheets are also rotated out of alignment (not shown). **b**, For small rotation angles, a 'moiré' pattern is produced in which the local stacking arrangement varies periodically. Cao *et al.*^{1,2} have observed that, for rotation angles of less than 1.05°, regions in which the atoms are directly above each other (the lighter regions in the pattern) form narrow electron energy bands, in which electron 'correlation' effects are enhanced. This results in the generation of a non-conducting state² (a Mott insulator), which can be converted into a superconducting state¹ if charge carriers are added to the graphene system.

effect completely dominates, and electron motion between layers is suppressed by a kinematic barrier⁵.

However, at very low rotation angles, a moiré pattern is produced by the misaligned lattices (Fig. 1b); the unit cell is greatly enlarged and so the effects of this come into play^{6,7}. The misalignment of the band structure essentially disappears, and theory predicts that the low-energy electronic states are completely reconstructed⁷. Coupling between electrons in the different layers becomes strong, and new narrow bands emerge at certain 'magic' rotation angles below 1.05° when the bilayer system is close to charge neutrality. Electrons in these narrow bands are found mainly in regions of the moiré pattern in which the atoms are stacked directly above each other (the light regions in Fig. 1b). In these circumstances, the bilayer can be thought of as a synthetic, triangular lattice of weakly coupled quantum dots (tiny semiconductor particles that bind electronic states) with a residual tunnelling of electrons between them⁶.

Cao *et al.* fabricated twisted bilayer graphene so that the sheets are rotated at magic angles, and accumulated or depleted charge carriers in the system to study how the charge-transport properties of the system depend on the filling of the energy bands. The authors observed² strong insulating behaviour when each unit cell of the synthetic lattice contained four charge carriers, a density that corresponds to complete filling of the bands. Intriguingly, they also find evidence for additional insulating states at lower densities in which the number of carriers per unit cell is an integer, but for which the narrow energy bands of the system are fractionally occupied. This suggests that the additional states are Mott insulating states, in which free motion of the carriers is prevented by their mutual repulsion, producing gridlock on the lattice. Mott insulators are a strongly correlated, non-conducting form of matter.

Even more intriguing is what happens when charge carriers are added to the Mott-insulator states associated with half-full unit cells of the synthetic lattice. The authors observe¹ that the system enters a state that has zero electrical resistance below a critical temperature of approximately 1.7 kelvin, in a phase change known as a Berezinskii–Kosterlitz–Thouless transition, thus forming a 2D superconductor. This transition temperature is remarkably high, given the very low carrier density achieved in these measurements (10¹¹ charge carriers per square centimetre). The high transition temperature and the apparent connection to correlated insulating states invites comparison of this superconducting state to that of a family of 'unconventional' superconductors⁸, which also have a close relationship with other strongly correlated electronic ground states. Twisted bilayer graphene might therefore be a useful experimental system in which to

investigate the mechanism of unconventional superconductivity.

In the meantime, Cao and colleagues' discoveries will stimulate a wave of activity as scientists seek to unwind the microscopic basis for the reported striking phenomena. The findings also demonstrate the promise of using twisted bilayer graphene as a flexible and tunable platform in which correlated electronic phenomena can be readily observed, and possibly even engineered and exploited⁹. ■

Eugene J. Mele is in the Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

STRUCTURAL BIOLOGY

Two-pore channels open up

Two-pore channels span the membranes of acidic organelles inside cells. A structural and functional analysis reveals secrets about how these channels open to allow ions to pass across the membrane. [SEE LETTER P.130](#)

SANDIP PATEL

Two-pore channels (TPCs) are an ancient family of ion channels that are unusual because they are found, not at the cell surface, but spanning the membranes of acidic organelles such as endosomes and lysosomes. These organelles mediate biomolecule transport and breakdown, and serve as stores of calcium ions¹ (Ca²⁺). TPCs are key for several organellar functions — releasing Ca²⁺ into the cytoplasm to control trafficking of material such as receptor proteins and viruses, for instance, and stabilizing junctions with other organelles^{1,2}. They are increasingly being associated with disorders such as Parkinson's disease, and are therefore emerging as potential therapeutic targets¹. Detailed structural information is scant, but advances in cryo-electron microscopy are revolutionizing our ability to study ion channels. On page 130, She *et al.*³ use this technique to provide the first detailed view of an animal TPC.

Previous work^{4,5} has reported the atomic structure of a plant TPC. This consists of two subunits, each containing two similar transmembrane domains (6-TMI and 6-TMII) connected by a large cytoplasmic linker. 6-TMI and 6-TMII are in turn each made up of six membrane-spanning regions, dubbed S1–S6. The pore through which ions flow is formed by S5 and S6 from each transmembrane domain in each subunit.

She *et al.* resolved the structure of mouse TPC1. Their results revealed that the overall

e-mail: mele@physics.upenn.edu

1. Cao, Y. *et al.* *Nature* **556**, 43–50 (2018).
2. Cao, Y. *et al.* *Nature* **556**, 80–84 (2018).
3. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. *Rev. Mod. Phys.* **81**, 109 (2009).
4. Zibrov, A. A. *et al.* *Nature* **549**, 360–364 (2017).
5. Lopes dos Santos, J. M. B., Peres, N. M. R. & Castro Neto, A. H. *Phys. Rev. Lett.* **99**, 256802 (2007).
6. de Laissardière, G. T., Mayou, D. & Magaud, L. *Phys. Rev. B* **86**, 125413 (2012).
7. Bistritzer, R. & MacDonald, A. H. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
8. Scalapino, D. J. *Rev. Mod. Phys.* **84**, 1383–1417 (2012).
9. Kim, K. *et al.* *Proc. Natl Acad. Sci. USA* **114**, 3364–3369 (2017).

This article was published online on 5 March 2018.

folding of this channel is, as expected, like that of plant TPC. Nonetheless, there is a surprising degree of structural conservation between the linkers, given that animal and plant TPCs have very different amino-acid sequences in this region.

There are some structural differences, however. In plant TPC, the linker binds Ca²⁺ to help open the channel^{4,5}. But Ca²⁺ binding by mouse TPC1 is unlikely, because amino acids essential for this interaction are missing. And the authors show that the carboxy-terminal domain of mouse TPC1, which is longer than the equivalent domain in plant TPC, forms a horseshoe-shaped arrangement of four helices that makes direct contact with the linker. This animal-specific feature probably serves to fine-tune channel activity.

Activation of animal TPCs is complex and multifaceted. These channels were originally identified^{6,7} as the targets for a messenger molecule called NAADP, which releases Ca²⁺ from acidic organelles⁸. Subsequent work revealed⁹ that TPCs are also activated by the lipid PI(3,5)P₂. In addition, TPC1 is regulated by changes in voltage across the organelle membrane^{10,11}. She *et al.* demonstrated that both PI(3,5)P₂ and voltage changes are required to open TPC1; neither alone is sufficient (they did not examine NAADP). The authors then resolved structures of TPC1 in both the absence and presence of PI(3,5)P₂, giving insight into the structural transitions that occur during channel opening. This analysis produced two key findings.

First, the group pinpointed the PI(3,5)P₂ binding site, which lies in 6-TMI (Fig. 1). Mutation of any one of several amino-acid residues in the network that forms this binding site can prevent TPC1 activation by PI(3,5)P₂. Interestingly, two of these residues — arginines in a short linker between S4 and S5 — are also required¹² for channel activation by NAADP. This suggests that PI(3,5)P₂ probably acts as a cofactor for NAADP action. Comparison of the free and PI(3,5)P₂-bound forms of TPC1 revealed that a single lysine residue in S6 transmits conformational changes to the pore in response to PI(3,5)P₂ binding, thus directing the first stage of channel opening.

Second, the authors found that changes in voltage are sensed by arginine residues in 6-TMII (Fig. 1). Both 6-TMI and 6-TMII contain sequences in S1–S4 that are reminiscent of voltage sensors in other channels, but only 6-TMII has a specific helix in S4 that is required for voltage gating. The 6-TMII voltage sensor is in an upward, ‘activated’ form in both structures obtained by the authors — in this form, it can probably transmit changes to the pore, to which it is adjacent, completing opening of the channel.

She and colleagues’ work on TPCs from animals, together with analyses^{4,5} of plant TPCs, indicate that both 6-TMI and 6-TMII cooperate to open the channel. 6-TMII is a target for voltage changes in both proteins. By contrast, 6-TMI is targeted directly by PI(3,5)P₂ in animal TPC1, and indirectly by Ca²⁺ in plant TPC. This is a prime example of how evolutionarily distant proteins have adapted to conserve a core function.

Which ions pass through animal TPCs

once they open? Much research suggests that these channels are non-selective, like plant TPC, but some work indicates that they are selective for sodium ions^{1,9} (Na⁺). She *et al.* found that TPC1 was about 70 times more permeable to Na⁺ than to potassium ions (K⁺). Their structures reveal that the narrowest part of the pore through which ions are filtered is shaped like an oblong ‘coin slot’, constricted by specific asparagine residues. The authors provide evidence that these residues allow the small Na⁺ ions through, but not the larger K⁺ ions. This sieve effect is unlikely to explain the authors’ data indicating that TPC1 apparently selects for Na⁺ over Ca²⁺, because these ions are about the same size. However, the electrophysiological experiments used by the researchers to determine ion selectivity were performed under very different conditions from those in live cells, where the permeability of TPCs to Ca²⁺ is readily demonstrable¹³.

In sum, She and colleagues’ structures provide major insight into how TPCs work. They join recently reported structures^{14–16} for a related family of ion channels, the TRP mucolipins (TRPMLs). Like TPCs, TRPMLs reside in acidic organelles, are activated by PI(3,5)P₂ and release Ca²⁺ to control cellular functions such as gene transcription¹⁷. The PI(3,5)P₂ binding site in TRPMLs is probably in the protein’s amino-terminal region^{16,17} and is thus very different from that in TPCs, although it has yet to be directly observed.

These rapid advances in the structural biology of organellar ion channels will aid future attempts to rationally design drugs that modulate ion flux through the channels. This is

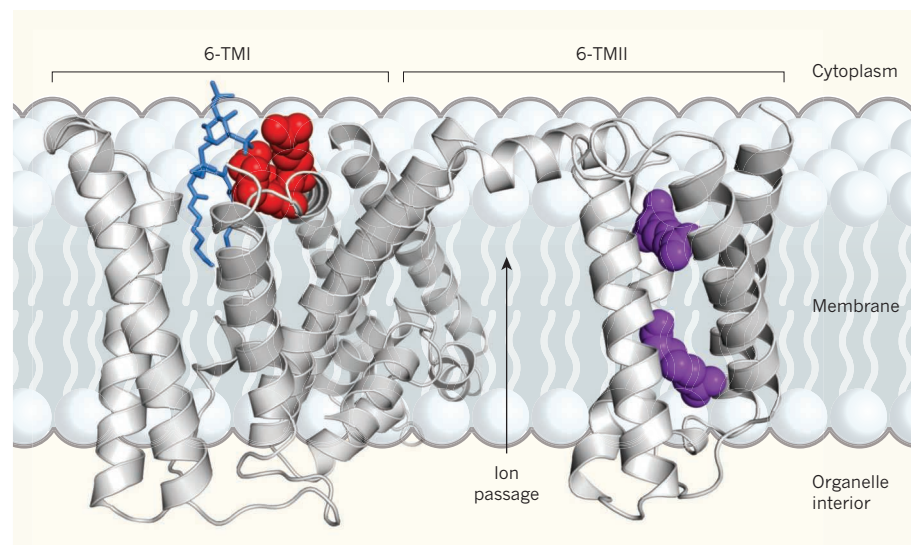


Figure 1 | Structure of mouse TPC1. She *et al.*³ have resolved the structure of the channel protein TPC1, which is found at organelle membranes. TPC1 has two subunits, each of which contains two transmembrane domains (6-TMI and 6-TMII), connected by a linker. Square brackets indicate the top of the helices that make up 6-TMI and 6-TMII. Here, only one subunit is depicted. The authors found that channel activation requires the lipid PI(3,5)P₂ (blue), which binds to arginine amino-acid residues (red) in 6-TMI, and voltage sensing through arginine residues (purple) in 6-TMII. Activation results in the flow of ions through the central pore region into the cytoplasm.



50 Years Ago

The University of Loughborough is already receiving encouraging response to its announcement last week of a new type of vacation course. Three weeks of courses are being arranged in July for technologists, scientists, managers, teachers and — here lies the novelty — for their spouses. Provision is being made for children so that families can be together while parents catch up on some mid-career training ... Twenty-four technical courses during the three weeks will cover such subjects as optics, ultrasonics ... statistics and management ... Cultural courses for spouses ... will cover industrial archaeology, music, drama and new techniques of food production ... Accommodation is provided on the campus for families at reasonable rates ... While hoping to provide all the facilities of a holiday camp, the university believes that its vacation courses will be more valuable than the description “intellectual Butlins” implies.

From *Nature* 6 April 1968

100 Years Ago

The recent development of aviation has provided a means of observing clouds which is much superior to any hitherto known. A modern aeroplane can reach the clouds in a very short time, and in many cases get above them. Observations of temperature can easily be obtained, and probably humidity observations would present no great difficulties. The “bumps” experienced also give some information as to the nature of the disturbance causing the formation of the clouds. It is well known that the two most important processes which cause clouds to form are (1) the mixture of layers of air of high humidity and different potential temperature, (2) adiabatic expansion due to upward movement.

From *Nature* 4 April 1918

pertinent as the number of diseases found to be associated with channel abnormalities grows. Mutations in TRPML1 cause a lysosomal storage disorder affecting children, and TPCs have been implicated in fatty liver disease, Ebola infection and several neurodegenerative disorders^{17,18}. In this context, a human TPC structure would be most welcome.

Another challenge is to resolve the structure of TPC2. This protein is regulated by NAADP and PI(3,5)P₂, but not by changes in voltage — begging the question of how conformational changes in one TM domain are transmitted to the other to allow channel opening. No

doubt, TPCs will reveal further secrets through forthcoming structures. ■

Sandip Patel is in the Department of Cell and Developmental Biology, University College London, London WC1E 6BT, UK.
e-mail: patel.s@ucl.ac.uk

1. Patel, S. *Sci. Signal.* **8**, re7 (2015).
2. Kilpatrick, B. S. *et al. Cell Rep.* **18**, 1636–1645 (2017).
3. She, J. *et al. Nature* **556**, 130–134 (2018).
4. Guo, J. *et al. Nature* **531**, 196–201 (2016).
5. Kintzer, A. F. & Stroud, R. M. *Nature* **531**, 258–264 (2016).
6. Brailoiu, E. *et al. J. Cell Biol.* **186**, 201–209 (2009).
7. Calcraff, P. J. *et al. Nature* **459**, 596–600 (2009).

8. Churchill, G. C. *et al. Cell* **111**, 703–708 (2002).
9. Wang, X. *et al. Cell* **151**, 372–383 (2012).
10. Rybalchenko, V. *et al. J. Biol. Chem.* **287**, 20407–20416 (2012).
11. Cang, C., Bekele, B. & Ren, D. *Nature Chem. Biol.* **10**, 463–469 (2014).
12. Patel, S., Churamani, D. & Brailoiu, E. *Cell Calcium* **68**, 1–4 (2017).
13. Ruas, M. *et al. EMBO J.* **34**, 1743–1758 (2015).
14. Schmiede, P., Fine, M., Blobel, G. & Li, X. *Nature* **550**, 366–370 (2017).
15. Hirschi, M. *et al. Nature* **550**, 411–414 (2017).
16. Chen, Q. *et al. Nature* **550**, 415–418 (2017).
17. Grimm, C., Butz, E., Chen, C.-C., Wahl-Schott, C. & Biel, M. *Cell Calcium* **67**, 148–155 (2017).
18. Patel, S. *Messenger* **5**(1–2), 24–29 (2016).

This article was published online on 21 March 2018.

MICROBIOLOGY

Bacterial persister cells tackled

Chronic infections can be hard to treat because slow-growing bacteria known as persister cells are usually unharmed by antibiotics. The identification of molecules that target such cells might provide a solution. [SEE LETTER P.103](#)

JULIAN G. HURDLE & ADITI DESHPANDE

The use of antibiotics to treat an infection can be unsuccessful when bacteria evade such drugs through genetic changes that endow them with antibiotic resistance. Pathogenic bacteria can also avoid antibiotic-mediated destruction through another route: some bacterial cells enter a metabolically inactive or dormant state to become persister cells, which grow slowly or not at all. Most antibiotics were discovered in experiments that tested the ability of compounds to

inhibit bacterial growth, and they are therefore often ineffective for treating non-growing persister cells¹. On page 103, Kim *et al.*² now report the identification of small molecules that can kill persister cells.

Persister cells are the source of many of the recurrent bacterial infections that affect people, for example those associated with implanted medical devices, such as the heart infection endocarditis, and also lung infections that can arise in cystic fibrosis^{1,3}. Curing such chronic infections can require surgery, which places an added health burden on patients.

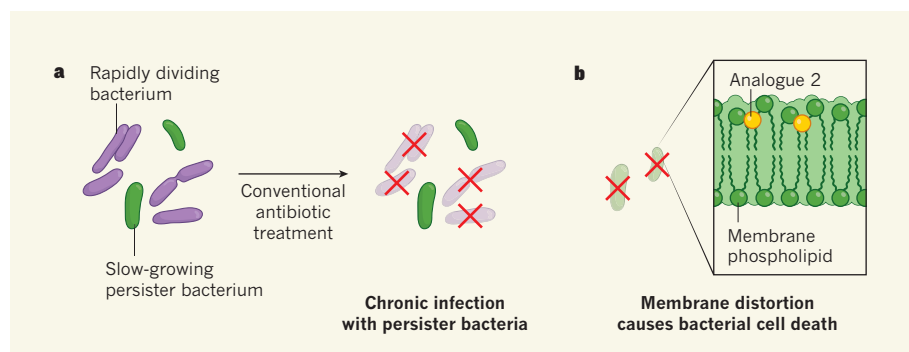


Figure 1 | A retinoid compound destroys bacterial persister cells. **a**, Bacterial populations commonly consist of rapidly dividing cells and slow-growing cells that are known as persister cells. When bacteria are treated with conventional antibiotics, the rapidly dividing cells are destroyed. However, the persister cells can remain, giving rise to chronic infection. **b**, Kim *et al.*² report a retinoid compound called analogue 2 that was optimized for targeting persister bacteria. In a mouse model of bacterial infection, the authors found that analogue 2 could kill persister cells. Electron microscopy and computer modelling revealed that analogue 2 probably binds to phospholipid molecules in the bacterial membrane, resulting in membrane distortion that might help to kill the bacteria.

And any necessary extended periods of treatment with antibiotics will increase the probability that bacteria evolve resistance. The development of treatments for killing persister cells is therefore urgently needed⁴, especially to target persisters that arise in infections with strains of the ‘superbug’ bacterium methicillin-resistant *Staphylococcus aureus* (MRSA), which is resistant to several common antibiotics. Infection with MRSA is associated with illness and death, particularly among people with invasive infections⁵.

Kim and colleagues decided to search for molecules that could offer protection from MRSA infection, using the roundworm *Caenorhabditis elegans* as a model system. Taking a high-throughput approach, the authors tested the ability of around 82,000 small synthetic molecules to protect worms from death mediated by MRSA infection. Of the 185 compounds that conferred protection, the authors focused on two molecules called CD437 and CD1530, both of which kill MRSA cells rapidly and can also target *Enterococcus faecium*, a bacterium that is linked to endocarditis. Unfortunately, these compounds had no effect against Gram-negative bacteria — a group that includes *Escherichia coli* — for which new therapeutic options are desperately needed because they, too, can form antibiotic-resistant superbugs.

CD437 and CD1530 belong to a class of molecule known as the retinoids, which are structurally similar to vitamin A. Since the 1960s, retinoids have been developed to treat various conditions, including acne⁶. Subsequent synthetic modification of the retinoids has therefore generated derivatives that are often present in chemical libraries used in drug discovery.

The authors concluded that prompt killing of MRSA cells occurred when the two retinoid molecules distorted the structure of the bacterial membrane’s lipid bilayer. Kim and colleagues then carried out electron-microscopy studies, which revealed that the retinoid treatment caused curvature and folding of the bacterial membrane but did not result in membrane destruction.

The bacterial membrane is a permeability

barrier that is essential for many cellular functions¹, and it contains proteins that are crucial to controlling the uptake of nutrients, the release of waste and the production of energy in the form of molecules of ATP. By distorting the membrane, the retinoid molecules probably affect the import and export of solutes, as well as several other essential cellular functions that rely on membrane integrity. However, a mode of antibacterial action that involves simply attacking the bacterial membrane is not guaranteed to kill persister cells. For example, the authors report that the membrane-targeting antibiotics nigericin and valinomycin do not kill MRSA persister cells.

Through computer simulations, the authors explored how the retinoid molecules might interact with the bacterial membrane. They determined that the polar side groups of CD437 and CD1530 could bind to the hydrophilic heads of phospholipids in the membrane, enabling the retinoid molecules to lodge in the lipid bilayer of a bacterium. Such simulations are a powerful tool that could be used to guide the optimization of antibiotics that can selectively attack the lipid bilayer of bacterial membranes without disrupting their mammalian counterparts and causing toxicity to patients.

A major concern is how to optimize small molecules such as the retinoids to enable such selectivity. Although the authors found that CD437 and CD1530 did not destroy the lipid membranes of human red blood cells, the

molecules were able to kill human liver-cancer cells grown *in vitro*, which is consistent with the previously reported anticancer properties of the retinoids⁶.

The authors generated structural variants of CD437, producing a compound they called analogue 2 that did not kill normal or cancerous human liver cells grown *in vitro*, but did retain the ability to kill MRSA persister cells (Fig. 1). In experiments in mice, analogue 2 remained in circulation in the animals' bodies

“Some bacterial cells enter a metabolically inactive or dormant state to become persister cells.”

for several hours at high enough concentrations to kill MRSA persister cells, but did not give rise to signs of toxicity such as liver or kidney damage. Remarkably, the authors showed in mice that analogue 2

could tackle what would generally be considered to be a treatment-resistant form of MRSA. This animal model mimics chronic infection with MRSA in immunocompromised people, for whom the prognosis is poor with conventional antibiotic treatments such as vancomycin because of the large number of MRSA persister cells that are present^{4,7}.

The authors found that the effects of analogue 2 on bacterial infections could be boosted by the presence of the antibiotic gentamicin, an inhibitor of bacterial protein synthesis that lacks activity against persister cells. It will be interesting to determine whether

MRSA persister cells respond to the retinoids by reactivating their cellular metabolism, thereby making them more susceptible to being killed by drugs such as gentamicin that would otherwise be ineffective.

Molecules such as analogue 2 might be suitable candidates for drugs that decrease the time required to successfully treat chronic infections that are characterized by high loads of dormant bacteria. In an era in which the development of antibiotics is struggling to keep pace with the spread of resistant bacteria, the identification of compounds such as analogue 2 could help researchers to win victories in the long fight against bacterial infectious diseases. ■

Julian G. Hurdle and Aditi Deshpande are in the Center for Infectious and Inflammatory Diseases, Institute of Biosciences and Technology, Texas A&M Health Science Center, Houston, Texas 77030, USA. e-mails: jhurdle@ibt.tamhsc.edu; adeshpande@medicine.tamhsc.edu

1. Hurdle, J. G., O'Neill, A. J., Chopra, I. & Lee, R. E. *Nature Rev. Microbiol.* **9**, 62–75 (2011).
2. Kim, W. *et al.* *Nature* **556**, 103–107 (2018).
3. Costerton, J. W., Stewart, P. S. & Greenberg, E. P. *Science* **284**, 1318–1322 (1999).
4. Lewis, K. *Nature Rev. Microbiol.* **5**, 48–56 (2007).
5. Dantes, R. *et al.* *JAMA Int. Med.* **173**, 1970–1978 (2013).
6. Álvarez, R., Vaz, B., Gronemeyer, H. & de Lera, Á. R. *Chem. Rev.* **114**, 1–125 (2014).
7. Conlon, B. P. *et al.* *Nature* **503**, 365–370 (2013).

This article was published online on 28 March 2018.

MATERIALS SCIENCE

Observations of the birth of crystals

Different forms of molecular crystals often have distinct properties, which can greatly influence their potential applications. A way of controlling the crystal form of a protein has now been reported. SEE LETTER P.89

ROBERT G. ALBERSTEIN & F. AKIF TEZCAN

The second law of thermodynamics dictates that all things tend towards disorder. Yet molecules and other microscopic particles in liquids frequently arrange themselves into perfectly ordered arrays — crystals — without violating this law. Moreover, a given molecule can often arrange itself into more than one type of array, producing different crystal forms known as polymorphs. These polymorphs can have remarkably different properties despite being composed of the same building blocks.

On page 89, Van Driessche *et al.*¹ report experimental observations of protein molecules as they begin to assemble into clusters that then evolve into distinct polymorphs. Their findings bring fresh insight to the important processes of crystal formation and growth, and polymorph selection.

The everyday consequences of crystal polymorphism are perhaps highlighted best by pharmaceutical drugs, most of which are administered as crystalline solids². Polymorphs of drug molecules often exhibit considerable variation in their ease of manufacture, their shelf life and — crucially — their

physical and chemical properties, which greatly influence their physiological efficacies². The selection and manufacture of appropriate polymorphs is a major and costly component of the drug-development process, yet processes for polymorph selection are largely conducted on the basis of trial and error, rather than through molecular design.

The development of rational approaches for the design and control of crystal growth, as well as for polymorph selection, requires an understanding of nucleation — the initial stages of crystallization, in which the building blocks begin to form clusters known as nuclei. Unfortunately, there are two main hurdles to capturing and characterizing crystals at birth. First, the nuclei are typically too small to be visualized in 3D space using most experimental methods, especially when they consist of atoms or small molecules. Second, such nuclei are, by definition, unstable and therefore form only transiently.

To address the first issue, Van Driessche and colleagues used the protein glucose isomerase (GI) as a building block, the box-like shape and nanometre dimensions of which make it relatively easy to identify using a technique called cryo-transmission electron microscopy (cryo-TEM). And to overcome the second issue, they

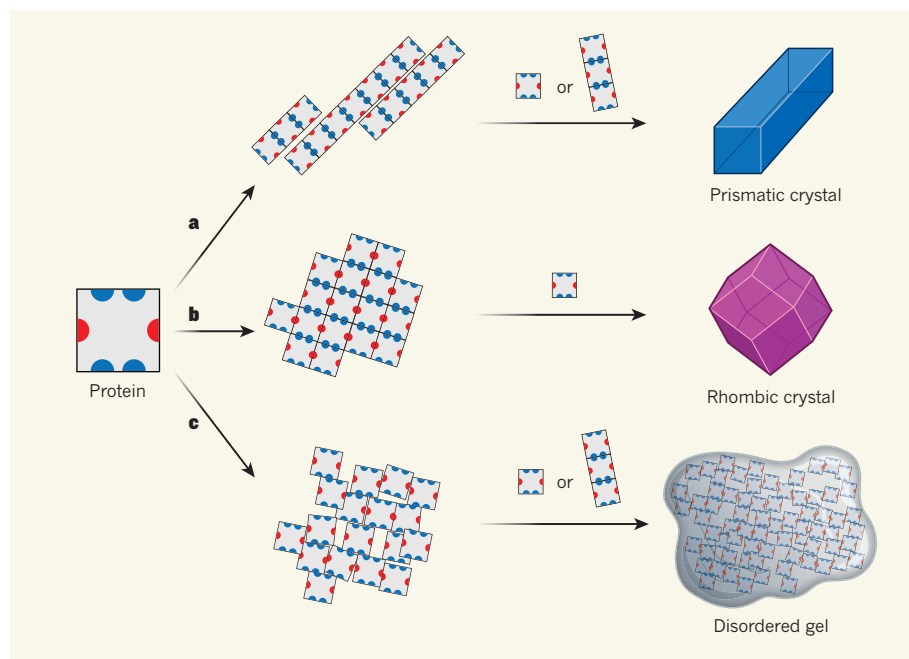


Figure 1 | Nucleation determines crystal growth and structure. Van Driessche *et al.*¹ have worked out how interactions between molecules of a box-shaped protein affect its nucleation (the formation of molecular clusters known as nuclei that act as ‘seeds’ for crystals) and, in turn, the growth and structure of crystals of the protein. Coloured semicircles represent sites of interaction on the molecules. **a**, Under conditions in which interactions between blue sites are stronger than those between red sites, rod-shaped nuclei form. Subsequent crystal growth occurs through the addition of either individual proteins or rods, and results in the formation of prismatic crystals. **b**, When the blue interactions are as strong as the red interactions, the nuclei have smaller aspect ratios than in **a** and grow equally in all directions through the addition of individual proteins. The resulting crystals are rhombic. **c**, Strong but indiscriminate interactions cause the molecules to form disordered clusters, which produce similarly disordered gels. (Adapted from a figure supplied by R. G. Alberstein.)

used a protocol in which protein samples were rapidly frozen to about -183°C , a temperature at which essentially all motion by GI molecules stops. This enabled them to take snapshots of the crystallization process at time intervals ranging from seconds to minutes. The authors thus captured the nucleation and growth of GI crystals with sufficiently high temporal and spatial resolution for them to work out how the emergent crystal morphologies depend on specific interactions formed between GI molecules.

To initiate crystallization, the authors mixed GI with ammonium sulfate or polyethylene glycol (PEG), which are commonly used as agents for modulating protein solubility and the strength of interactions between proteins. They observed that, at high concentrations of ammonium sulfate, GI molecules rapidly line up into rods that are a few proteins in length. These rods then align side by side, while also growing longer, to yield macroscopic crystals with a rectangular, prism-like shape that resembles the rod-shaped nuclei (Fig. 1a). By contrast, low concentrations of PEG cause the GI nuclei to grow more slowly and evenly in all dimensions, to yield rhombic crystals with a diamond-like shape (Fig. 1b). And at high concentrations of PEG, the GI molecules become locked into structures that are best described as disordered gels rather than crystals (Fig. 1c).

Van Driessche and colleagues’ cryo-TEM images of the GI nuclei are detailed enough to be compared with known 3D atomic structures of GI crystals. Such comparisons enabled the authors to propose plausible models for the arrangement of GI molecules in the nuclei, as well as for the specific interactions between the molecules. The authors then designed mutants of GI in which a key amino-acid residue at each interaction site in the various nuclei was replaced with another residue. The mutant proteins were unable to form their corresponding polymorph and either produced the alternative crystal polymorph or aggregated into gels, depending on the conditions. These observations validated the proposed structural models and demonstrated that polymorphs could be selected predictably.

Classical nucleation theory (CNT) posits that crystallization must start with the formation of a nucleus that has the same molecular order and arrangement as do the macroscopic crystals, and that the building blocks are added one by one to the nucleus³. In the past two to three decades, CNT has been largely superseded by two-step or multistep nucleation models in which an amorphous, high-density liquid phase forms, and then transitions into either a single crystalline domain (as occurs in some inorganic nucleation processes⁴) or numerous small, locally ordered clusters,

which align to form growing crystals by a process known as oriented attachment. In light of these competing views, characterization of the nucleation events for various systems has been the subject of much experimental and theoretical work³.

Van Driessche and co-workers’ observations, particularly of the prismatic GI crystals, indicate that elements of both CNT and multistep nucleation might be at play. The authors uncover no evidence of an amorphous liquid phase, and the smallest GI rods, which they captured at very early time points (just 20 seconds after the addition of ammonium sulfate), have the same crystalline registry as the mature crystals — findings that are in accord with CNT. However, the rods undergo oriented attachment during crystal maturation, as in the multistep model. The picture is less clear for the rhombic GI crystals, the precursors of which become observable only after several minutes — at which point they are already bigger than a nucleus. It could be that an amorphous, high-density liquid phase does form in this case but escapes detection because of its instability or the low contrast of the cryo-TEM images.

We can, however, be certain that rapidly advancing techniques such as cryo-TEM, liquid-cell transmission electron microscopy^{5,6} and *in situ* atomic force microscopy⁷, complemented by theory and computational modelling⁸, will continue to provide intriguing results with which to refine our understanding of crystal nucleation and growth. More practically, Van Driessche and colleagues’ study shows that the crystallization or self-assembly pathways of proteins can be rationally engineered at the molecular level to obtain a desired polymorph. This feat is particularly notable, given that the number of protein-based agents being used as pharmaceutical drugs is increasing⁹, and that synthetic protein assemblies and crystals are being designed and constructed to have unusual and potentially useful properties¹⁰. ■

Robert G. Alberstein and F. Akif Tezcan are in the Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, USA. e-mail: tezcan@ucsd.edu

1. Van Driessche, A. E. S. *et al.* *Nature* **556**, 89–94 (2018).
2. Lee, A. Y., Erdemir, D. & Myerson, A. S. *Annu. Rev. Chem. Biomol. Eng.* **2**, 259–280 (2011).
3. De Yoreo, J. J. *et al.* *Science* **349**, aaa6760 (2015).
4. Pouget, E. M. *et al.* *Science* **323**, 1455–1458 (2009).
5. Yamazaki, Y. *et al.* *Proc. Natl Acad. Sci. USA* **114**, 2154–2159 (2017).
6. Parent, L. R. *et al.* *J. Am. Chem. Soc.* **139**, 17140–17151 (2017).
7. Chung, S., Shin, S.-H., Bertozzi, C. R. & De Yoreo, J. J. *Proc. Natl Acad. Sci. USA* **107**, 16536–16541 (2010).
8. Whitelam, S. *Phys. Rev. Lett.* **105**, 088102 (2010).
9. Walsh, G. *Nature Biotechnol.* **28**, 917–924 (2010).
10. Suzuki, Y. *et al.* *Nature* **533**, 369–373 (2016).

Unconventional superconductivity in magic-angle graphene superlattices

Yuan Cao¹, Valla Fatemi¹, Shiang Fang², Kenji Watanabe³, Takashi Taniguchi³, Efthimios Kaxiras^{2,4} & Pablo Jarillo-Herrero¹

The behaviour of strongly correlated materials, and in particular unconventional superconductors, has been studied extensively for decades, but is still not well understood. This lack of theoretical understanding has motivated the development of experimental techniques for studying such behaviour, such as using ultracold atom lattices to simulate quantum materials. Here we report the realization of intrinsic unconventional superconductivity—which cannot be explained by weak electron–phonon interactions—in a two-dimensional superlattice created by stacking two sheets of graphene that are twisted relative to each other by a small angle. For twist angles of about 1.1° —the first ‘magic’ angle—the electronic band structure of this ‘twisted bilayer graphene’ exhibits flat bands near zero Fermi energy, resulting in correlated insulating states at half-filling. Upon electrostatic doping of the material away from these correlated insulating states, we observe tunable zero-resistance states with a critical temperature of up to 1.7 kelvin. The temperature–carrier-density phase diagram of twisted bilayer graphene is similar to that of copper oxides (or cuprates), and includes dome-shaped regions that correspond to superconductivity. Moreover, quantum oscillations in the longitudinal resistance of the material indicate the presence of small Fermi surfaces near the correlated insulating states, in analogy with underdoped cuprates. The relatively high superconducting critical temperature of twisted bilayer graphene, given such a small Fermi surface (which corresponds to a carrier density of about 10^{11} per square centimetre), puts it among the superconductors with the strongest pairing strength between electrons. Twisted bilayer graphene is a precisely tunable, purely carbon-based, two-dimensional superconductor. It is therefore an ideal material for investigations of strongly correlated phenomena, which could lead to insights into the physics of high-critical-temperature superconductors and quantum spin liquids.

Strong interactions among particles lead to fascinating states of matter, such as quark–gluon plasmas, various forms of nuclear matter within neutron stars, strange metals and fractional quantum Hall states^{1–3}. An intriguing class of strongly correlated materials is the unconventional superconductors, which includes materials with a range of superconducting critical temperatures T_c , from heavy-fermion and organic superconductors with relatively low T_c (a few to a few tens of kelvin) to iron pnictides and cuprates that can have $T_c > 100$ K (refs 4–8). Despite extensive experimental efforts to characterize these materials, unconventional superconductors are challenging to study theoretically because the models that are typically used to describe them cannot be solved exactly, motivating the development of alternative approaches for investigating and modelling strongly correlated systems. One approach is to simulate quantum materials with ultracold atoms trapped in optical lattices, although technical advances are necessary to realize *d*-wave superfluidity with ultracold atoms at lower temperatures than are currently possible^{9,10}.

Here we report the observation of unconventional superconductivity in a two-dimensional superlattice made from graphene—specifically, ‘magic angle’ twisted bilayer graphene (TBG). Created by the moiré pattern between the two graphene sheets, the magic-angle TBG superlattice has a periodicity of about 13 nm, between that of crystalline superconductors (a few ångström) and optical lattices (about a micrometre). One of the key advantages of this system is the *in situ* electrical tunability of the charge carrier density in a flat band with a bandwidth of the order of 10 meV. This tunability enables us to study the phase diagram of unconventional superconductivity in unprecedented resolution, without relying on multiple devices that are possibly hampered by

different disorder realizations. The superconductivity that we observe has several features similar to that of cuprates, including dome structures in the phase diagram and quantum oscillations that point to small Fermi surfaces near a correlated insulator state. Furthermore, it occurs for record-low carrier densities of the order of 10^{11} cm^{-2} , orders of magnitude lower than the carrier densities of typical two-dimensional superconductors. The relatively high $T_c = 1.7$ K for such small densities puts magic-angle TBG among the superconductors with the strongest coupling, in the same league as cuprates and the recently identified FeSe thin layers¹¹. Our results establish magic-angle TBG as a purely carbon-based two-dimensional superconductor and, more importantly, as a relatively simple and highly tunable material that enables thorough investigation of strongly correlated physics.

Monolayer graphene has a linear energy dispersion at its charge neutrality point. When two aligned graphene sheets are stacked, the hybridization of their bands due to interlayer hopping results in fundamental modifications to the low-energy band structure depending on the stacking order (AA or AB). If an additional twist angle is present between layers, a hexagonal moiré pattern consisting of alternating AA- and AB-stacked regions emerges and acts as a superlattice modulation^{12–16}. The superlattice potential folds the band structure into the mini Brillouin zone. Hybridization between adjacent Dirac cones in the mini Brillouin zone has an effect on the Fermi velocity at the charge neutrality point, reducing it from the typical value^{12–18} of 10^6 m s^{-1} . At low twist angles, each electronic band in the mini Brillouin zone has a four-fold degeneracy of spins and valleys, the latter inherited from the original electronic structure of graphene^{12,17,19}. For convenience, we define the superlattice density $n_s = 4/A$ to be the density that

¹Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. ³National Institute for Materials Science, Namiki 1-1, Tsukuba, Ibaraki 305-0044, Japan. ⁴John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

corresponds to full-filling of each set of degenerate superlattice bands, where $A \approx \sqrt{3}a^2/(2\theta^2)$ is the area of the moiré unit cell, $a = 0.246$ nm is the lattice constant of the underlying graphene lattice and θ is the twist angle. In Supplementary Video, we present an animation of the way in which the band structure in the mini Brillouin zone of TBG evolves as the twist angle varies from $\theta = 3^\circ$ to $\theta = 0.8^\circ$, calculated using a continuum model for one valley¹².

Special angles, namely the ‘magic angles’, exist, at which the Fermi velocity drops to zero; the first magic angle is predicted¹² to be $\theta_{\text{magic}}^{(1)} \approx 1.1^\circ$. Near this twist angle, the energy bands near charge neutrality, which are separated from other bands by single-particle gaps, become remarkably flat. The typical energy scale for the entire bandwidth is about 5–10 meV (Fig. 1c)^{12,18}. Experimentally confirmed consequences of the flatness of these bands are high effective mass in the flat bands (as observed in quantum oscillations) and correlated insulating states at half-filling of these bands, corresponding to $n = \pm n_s/2$, where $n = CV_g/e$ is the carrier density defined by the gate voltage V_g (C is the gate capacitance per unit area and e is the electron charge)¹⁸. These insulating states are a result of the competition between Coulomb

energy and quantum kinetic energy, which gives rise to a correlated insulator at half-filling that has characteristics consistent with Mott-like insulator behaviour¹⁸. The doping density that is required to reach the Mott-like insulating states is $n_s/2 \approx (1.2\text{--}1.6) \times 10^{12} \text{ cm}^{-2}$, depending on the exact twist angle. Here we report transport data that clearly demonstrate that superconductivity is achieved as the material is doped slightly away from the Mott-like insulating state in magic-angle TBG. We observed superconductivity across multiple devices with slightly different twist angles, with the highest critical temperature that we achieved being 1.7 K.

Superconductivity in magic-angle TBG

In Fig. 1a we show the typical device structure of fully encapsulated TBG devices. The two sheets of graphene originate from the same exfoliated flake, which permits a relative twist angle that is controlled precisely to within about $0.1^\circ\text{--}0.2^\circ$ (refs 17, 20, 21). The encapsulated TBG stack is etched into a ‘Hall’ bar and contacted from the edges²². Electrical contacts are made from non-superconducting materials (thermally evaporated Au on a Cr sticking layer) to avoid any potential

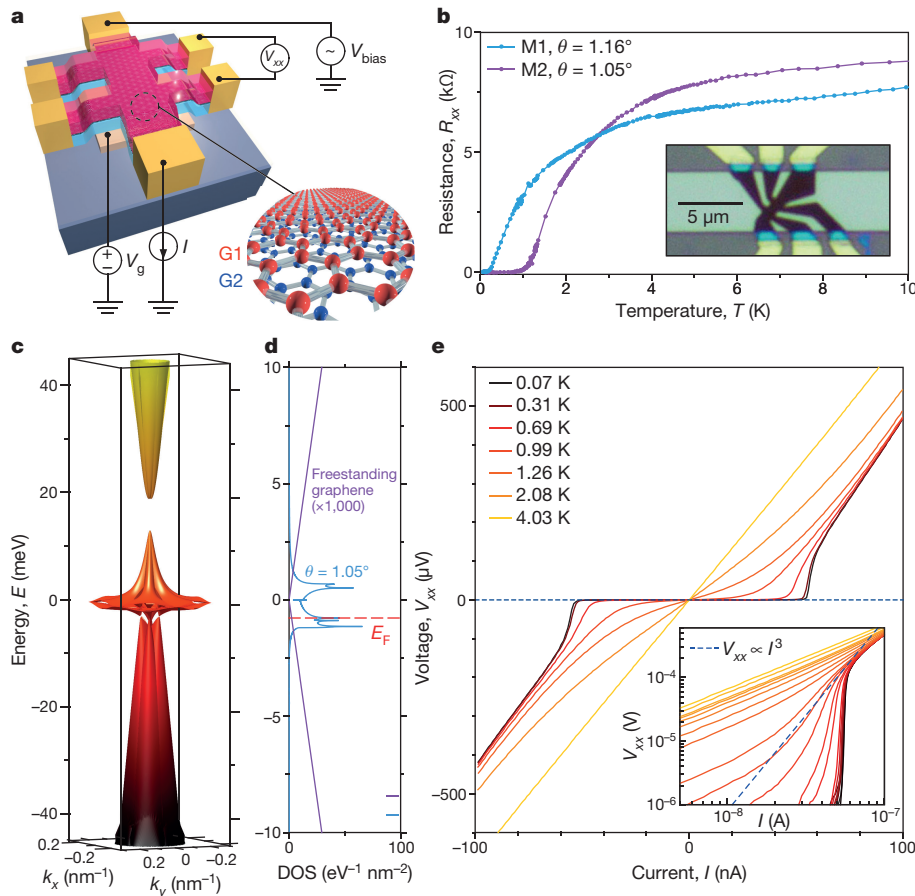


Figure 1 | Two-dimensional superconductivity in a graphene superlattice. **a**, Schematic of a typical twisted bilayer graphene (TBG) device and the four-probe (V_{xx} , V_g , I and the bias voltage V_{bias}) measurement scheme. The stack consists of hexagonal boron nitride on the top and bottom, with two graphene bilayers (G1, G2) twisted relative to each other in between. The electron density is tuned by a metal gate beneath the bottom hexagonal boron nitride layer. **b**, Four-probe resistance $R_{xx} = V_{xx}/I$ (V_{xx} and I are defined in **a**) measured in two devices M1 and M2, which have twist angles of $\theta = 1.16^\circ$ and $\theta = 1.05^\circ$, respectively. The inset shows an optical image of device M1, including the main ‘Hall’ bar (dark brown), electrical contact (gold), back gate (light green) and SiO_2/Si substrate (dark grey). **c**, The band energy E of TBG at $\theta = 1.05^\circ$ in the first mini Brillouin zone of the superlattice. The bands near charge neutrality ($E = 0$) have energies of less than 15 meV.

d, The DOS corresponding to the bands shown in **c**, for energies of -10 to $+10$ meV (blue; $\theta = 1.05^\circ$). For comparison, the purple lines show the total DOS of two sheets of freestanding graphene without interlayer interaction (multiplied by 10^3). The red dashed line shows the Fermi energy E_F at half-filling of the lower branch ($E < 0$) of the flat bands, which corresponds to a density of $n = -n_s/2$, where n_s is the superlattice density (defined in the main text). The superconductivity is observed near this half-filled state. **e**, Current–voltage (V_{xx} – I) curves for device M2 measured at $n = -1.44 \times 10^{12} \text{ cm}^{-2}$ and various temperatures. At the lowest temperature of 70 mK, the curves indicate a critical current of approximately 50 nA. The inset shows the same data on a logarithmic scale, which is typically used to extract the Berezinskii–Kosterlitz–Thouless transition temperature ($T_{\text{BKT}} = 1.0$ K in this case), by fitting to a $V_{xx} \propto I^3$ power law (blue dashed line).

proximity effects. The carrier density n is tuned by applying a voltage to a Pd/Au bottom gate electrode. In Fig. 1b we show the longitudinal resistance R_{xx} as a function of temperature for two magic-angle devices, M1 and M2, with twist angles of 1.16° and 1.05° , respectively. At the lowest temperature studied of 70 mK, both devices show zero resistance, and therefore a superconducting state. The critical temperature T_c as calculated using a resistance of 50% of the ‘normal’-state (non-superconducting) value is approximately 1.7 K and 0.5 K for the two devices that we studied in detail. In Fig. 1c, d we show a single-particle band structure and density of states (DOS) near the charge neutrality point calculated for $\theta = 1.05^\circ$. The superconductivity in both devices occurs when the Fermi energy E_F is tuned away from charge neutrality ($E_F = 0$) to be near half-filling of the lower flat band ($E_F < 0$, as indicated in Fig. 1d). The DOS within the energy scale of the flat bands is more than three orders of magnitudes higher than that of two uncoupled graphene sheets, owing to the reduction of the Fermi velocity and the increase in localization that occurs near the magic angle. However, the energy at which the DOS peaks does not generally coincide with the density that is required to half-fill the bands. In addition, we did not observe any appreciable superconductivity when the Fermi energy was tuned into the flat conduction bands ($E_F > 0$). In Fig. 1e we show the current–voltage (I – V_{xx} , where V_{xx} is the four-probe voltage, as defined in Fig. 1a) curves of device M2 at different temperatures. We observe typical behaviour for a two-dimensional superconductor. The inset shows a tentative fit of the same data to a $V_{xx} \propto I^3$ power law, as is predicted in a Berezinskii–Kosterlitz–Thouless transition in two-dimensional superconductors²³. This analysis yields a Berezinskii–Kosterlitz–Thouless transition temperature of $T_{BKT} \approx 1.0$ K at $n = -1.44 \times 10^{12} \text{ cm}^{-2}$, where, as before,

n is the carrier density induced by the gate and measured from the charge neutrality point (which is different from the actual carrier density involved in transport, as we show below).

In contrast to other known two-dimensional and layered superconductors, the superconductivity in magic-angle TBG requires the application of only a small gate voltage, corresponding to a minimal density of only $1.2 \times 10^{12} \text{ cm}^{-2}$ from charge neutrality, an order of magnitude lower than the value of $1.5 \times 10^{13} \text{ cm}^{-2}$ in $\text{LaAlO}_3/\text{SrTiO}_3$ interfaces and of $7 \times 10^{13} \text{ cm}^{-2}$ in electrochemically doped MoS_2 , among others²⁴. Therefore, gate-tunable superconductivity can be realized in a high-mobility system without the need for ionic-liquid gating or chemical doping. In Fig. 2a we show the two-probe conductance of device M1 versus n at zero magnetic field and at a 0.4-T perpendicular magnetic field. Near the charge neutrality point ($n = 0$), a typical V-shaped conductance is observed, which originates from the renormalized Dirac cones of the TBG band structure. The insulating states centred at approximately $\pm 3.2 \times 10^{12} \text{ cm}^{-2}$ (which corresponds to n_s for $\theta = 1.16^\circ$) are due to single-particle bandgaps in the band structure that correspond to filling ± 4 electrons in each superlattice unit cell. In between, there are conductance minima at ± 2 and ± 3 electrons per unit cell. These minima are associated with many-body gaps induced by the competition between the Coulomb energy and the reduced kinetic energy due to confinement of the electronic state in the superlattice near the magic angle; these gaps give rise to insulating behaviour near the integer fillings¹⁸. One possible mechanism for the gaps is similar to the gap mechanism in Mott insulators, but with an extra two-fold degeneracy (for the case of ± 2 electrons) from the valleys in the original graphene Brillouin zone^{17,18,25,26}. In the vicinity of -2 electrons per unit cell ($n \approx -1.3 \times 10^{12} \text{ cm}^{-2}$ to $n \approx -1.9 \times 10^{12} \text{ cm}^{-2}$) and at a

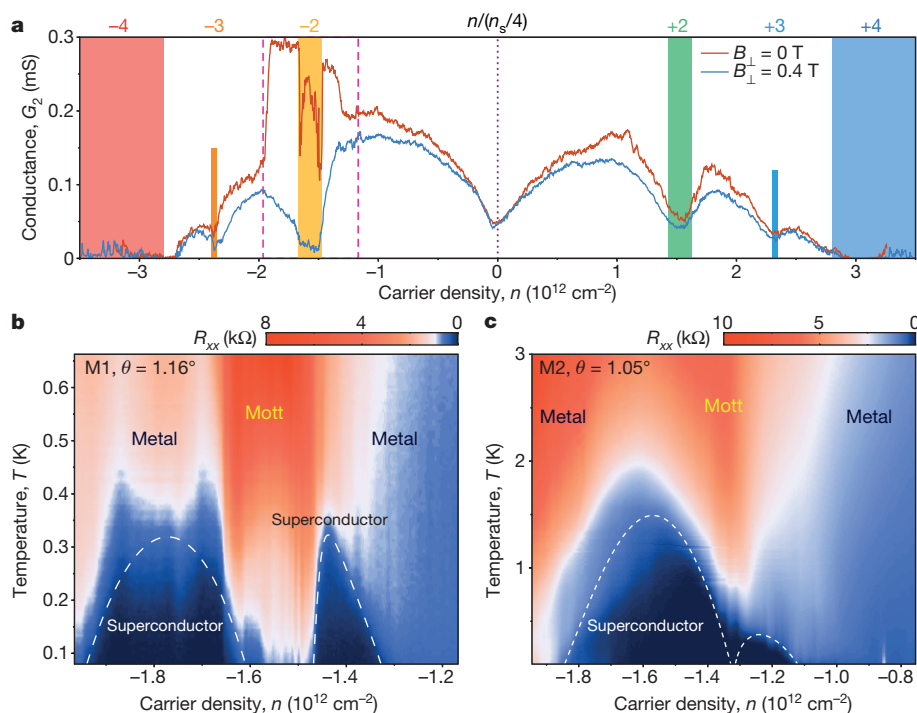


Figure 2 | Gate-tunable superconductivity in magic-angle TBG.

a, Two-probe conductance $G_2 = I/V_{\text{bias}}$ of device M1 ($\theta = 1.16^\circ$) measured in zero magnetic field (red) and at a perpendicular field of $B_\perp = 0.4 \text{ T}$ (blue). The curves exhibit the typical V-shaped conductance near charge neutrality ($n = 0$, vertical purple dotted line) and insulating states at the superlattice bandgaps $n = \pm n_s$, which correspond to filling ± 4 electrons in each moiré unit cell (blue and red bars). They also exhibit reduced conductance at intermediate integer fillings of the superlattice owing to Coulomb interactions (other coloured bars). Near a filling of -2 electrons per unit cell, there is considerable conductance enhancement at zero field that is suppressed in $B_\perp = 0.4 \text{ T}$. This enhancement signals the onset of

superconductivity. Measurements were conducted at 70 mK; $V_{\text{bias}} = 10 \mu\text{V}$.

b, Four-probe resistance R_{xx} , measured at densities corresponding to the region bounded by pink dashed lines in **a**, versus temperature. Two superconducting domes are observed next to the half-filling state, which is labelled ‘Mott’ and centred around $-n_s/2 = -1.58 \times 10^{12} \text{ cm}^{-2}$. The remaining regions in the diagram are labelled as ‘metal’ owing to the metallic temperature dependence. The highest critical temperature observed in device M1 is $T_c = 0.5 \text{ K}$ (at 50% of the normal-state resistance). **c**, As in **b**, but for device M2, showing two asymmetric and overlapping domes. The highest critical temperature in this device is $T_c = 1.7 \text{ K}$.

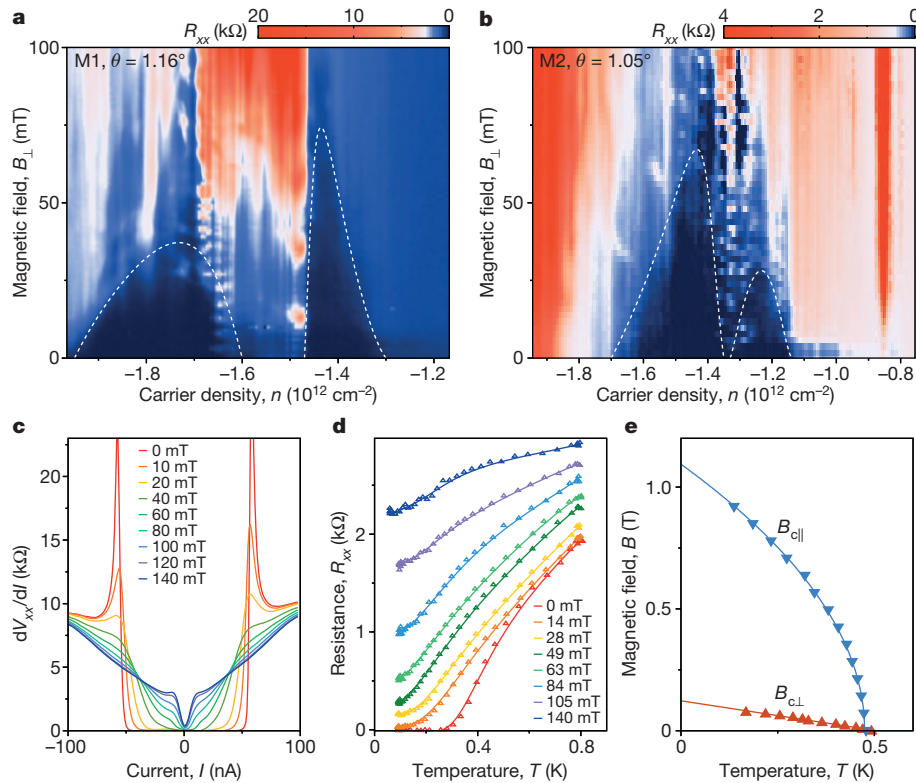


Figure 3 | Magnetic-field response of the superconducting states in magic-angle TBG. **a, b**, Four-probe resistance as a function of density n and perpendicular magnetic field B_{\perp} in devices M1 (**a**) and M2 (**b**). As well as the dome structures around half-filling (similar to those in Fig. 2b, c), there are oscillatory features near the boundary between the superconducting phase and the correlated insulator phase. These oscillations are indicative of phase-coherent transport through inhomogeneous regions in the device (Methods, Extended Data Fig. 1).

temperature of 70 mK, the conductance is substantially higher at zero magnetic field than it is in a perpendicular magnetic field of $B_{\perp} = 0.4$ T, consistent with mean-field suppression of a superconducting state by the magnetic field. Here, the maximum conductance is limited only by the contact resistance (Fig. 2a), which is absent in the four-probe measurements shown in the other figures.

In Fig. 2b, c we show the four-probe resistance of devices M1 and M2, respectively, as a function of density n and temperature T . Both devices show two pronounced superconducting domes on each side of the half-filling correlated insulating state. These features are similar to those associated with high-temperature superconductivity in cuprate materials. At the base temperature, the resistance inside the domes is lower than our measurement noise floor, which is more than two and three orders of magnitude lower than the normal-state resistance for devices M1 and M2, respectively. The I - V curves inside the domes display critical current behaviour (Fig. 1e), while being ohmic in the metallic phases outside the domes. Upon cooling while n is fixed at the middle of the half-filling state, the correlated insulating phase is exhibited at intermediate temperatures (from 1 K to 4 K); at lower temperatures, both devices exhibit signs of superconductivity at the lowest temperatures. Device M1 becomes weakly superconducting, whereas device M2 becomes fully superconducting. This behaviour may be explained by a coexistence of superconducting and insulating phases due to sample inhomogeneity.

Magnetic-field response

The application of a perpendicular magnetic field B_{\perp} to a two-dimensional superconductor creates vortices that introduce dissipation and gradually suppress superconductivity²³. In Fig. 3a, b we show the

resistance of devices M1 and M2 as a function of density and B_{\perp} . Both devices exhibit a maximum critical field of approximately 70 mT. The critical field varies strongly with doping density, showing two similar domes on each side of the half-filling state. Near the Mott-like insulating state ($n \approx -1.47 \times 10^{12} \text{ cm}^{-2}$ to $n \approx -1.67 \times 10^{12} \text{ cm}^{-2}$ for M1; $n \approx -1.25 \times 10^{12} \text{ cm}^{-2}$ to $n \approx -1.35 \times 10^{12} \text{ cm}^{-2}$ for M2), periodic oscillations of the resistance and critical current as a function of B_{\perp} appear (see Methods and Extended Data Fig. 1 for detailed analysis). The oscillations seem to originate from phase-coherent transport through arrays of Josephson junctions, similarly to superconducting quantum interference device (SQUID)-like superconductor rings around one or more insulating islands. These junction regions could be due to slight density inhomogeneities in the devices, which would cause a few islands to be doped into the insulating phase while other parts of the device remain superconducting. Apart from these oscillatory behaviours near the boundary of the half-filling insulating state, the critical current and zero resistivity inside the domes are gradually suppressed by B_{\perp} (Fig. 3c, d).

In Fig. 3e we show the critical magnetic field versus temperature for device M1, under perpendicular and parallel field configurations. The temperature dependence of the perpendicular critical field $B_{c\perp}$ is well described by Ginzburg–Landau theory: $B_{c\perp} = [\Phi_0/(2\pi\xi_{GL}^2)](1 - T/T_c)$, where $\Phi_0 = h/(2e)$ is the superconducting flux quantum, h is the Planck constant, and ξ_{GL} is the Ginzburg–Landau superconducting coherence length, determined from the fit to be $\xi_{GL} \approx 52 \text{ nm}$ at $T = 0$. On the other hand, the in-plane critical field dependence is not well explained by the Ginzburg–Landau theory for thin-film superconductors, owing to the atomic thickness of TBG (0.6 nm); at this thickness, the theory predicts an in-plane critical field of $B_{c\parallel} \geq 36 \text{ T}$ as the temperature

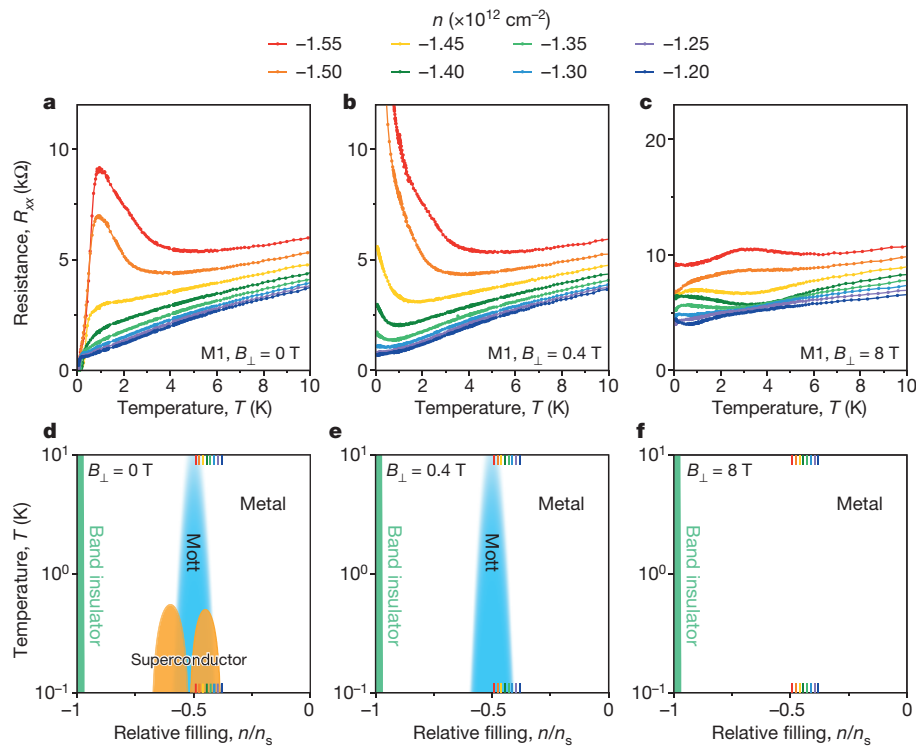


Figure 4 | Temperature–density phase diagrams of magic-angle TBG at different magnetic fields. **a–c**, R_{xx} – T curves for device M1 at different densities (see legend), measured in $B_{\perp} = 0$ T (**a**), $B_{\perp} = 0.4$ T (**b**) and $B_{\perp} = 8$ T (**c**). The magnetic field induces a superconductor–insulator–

metal transition at the lowest temperatures. **d–f**, Schematic phase diagrams for the magnetic fields in **a–c**. The horizontal axis shows the relative filling n/n_s . Short coloured lines at the top and bottom of the plots denote the densities plotted in **a–c**.

approaches zero²³. Instead, we interpret the dependence of T_c on the in-plane magnetic field B_{\parallel} as a result of paramagnetic pair-breaking owing to the Zeeman energy. The zero-temperature in-plane critical field is extrapolated to be around 1.1 T, which is higher than but close to the value in the Pauli limit of $B_p \approx 1.85$ T $K^{-1} \times T_c \approx 0.93$ T, estimated on the basis of the Bardeen–Cooper–Schrieffer (BCS) gap formula $\Delta \approx 1.76 k_B T_c$, where k_B is the Boltzmann constant.

We note that the superconductor–metal transition in magic-angle TBG is not sharp, so extracting both B_c and T_c has some uncertainty. Qualitatively, the dependence of the in-plane critical field on temperature is $B_{c\parallel} \propto (1 - T/T_c)^{1/2}$ near T_c (ref. 27). The results described above are consistent with the existence of two-dimensional superconductivity confined in an atomically thin space. As we show in the following, the coherence length ξ is comparable to the inter-particle spacing and might suggest that the system is driven close to a crossover between a BCS-like state and a Bose–Einstein condensate (the BCS–BEC crossover).

Phase diagram of magic-angle TBG

The phase diagram of magic-angle TBG consists of correlated insulator phases and superconducting phases, which can be realized via continuous tuning of temperature, magnetic field and carrier density. Similarly to the superconducting domes discussed above, the correlated Mott-like insulator phase at half-filling also assumes a dome shape, with a transition to a metallic phase at about 4–6 K and centred around half-filling density. It has been shown¹⁸ that the Mott-like insulator phase crosses over to a metallic phase upon application of a strong magnetic field of around 6 T either perpendicular or parallel to the devices. A plausible explanation for this crossover is that the many-body charge gap is closed by the Zeeman energy.

In Fig. 4a–c we show the resistance versus temperature data measured in device M1 at zero magnetic field, $B_{\perp} = 0.4$ T and $B_{\perp} = 8$ T, respectively. At zero field, we observe the transition from a metal at high temperatures (above 5 K) to a superconductor. Close to half-filling there

is an intermediate region in which insulating temperature dependence is observed from about 1 K to 4 K (above T_c); we identify this region as corresponding to the Mott-like insulating phase at half-filling. In a small magnetic field $B_{\perp} = 0.4$ T, which is above the critical magnetic field, the system remains an insulator down to zero temperature near half-filling and a metal away from half-filling. Finally, in a strong magnetic field $B_{\perp} = 8$ T, the correlated insulator phase is fully suppressed by the Zeeman effect and the system is metallic everywhere between $n = -n_s$ and the charge neutrality point. Our data highlight the rich phase space of metal–insulator–superconducting physics in magic-angle TBG²⁸. A schematic of the evolution of the phase diagram as the magnetic field increases is shown in Fig. 4d–f.

Quantum oscillations in the normal state

We studied quantum oscillations in the entire accessible density range, including in the vicinity of the correlated insulating state at which superconductivity occurs. In Fig. 5a, b we show the Shubnikov–de Haas oscillations in longitudinal resistance R_{xx} as a function of carrier density for the hole-doped region ($E_F < 0$) for device M2. The Landau levels in a TBG superlattice typically follow $n/n_s = N\phi/\phi_0 + s$, where $\phi = B_{\perp}A$ is the magnetic flux that penetrates each unit cell, $\phi_0 = h/e$ is the (non-superconducting) flux quantum, $N = \pm 1, \pm 2, \pm 3, \dots$ is the Landau-level index, $s = 0$ denotes the Landau fan that emanates from the Dirac point, and $s = \pm 1$ denote the Landau fans that result from electron-like or hole-like quasiparticles near the band edges of the single-particle superlattice bands in the mini Brillouin zone, which emanate from $\pm n_s$. The Landau levels also exhibit a four-fold degeneracy due to spins and valleys, and so the filling-factor sequence is $\pm 4, \pm 8, \pm 12, \dots$

Unexpectedly, in addition to these expected Landau fans, we also observe a Landau fan that emanates from the correlated insulating state at $-n_s/2$. This Landau fan has $N = -1/2, -1, -3/2, -2, \dots$ (that is, filling factors of $-2, -4, -6, -8, \dots$) and $s = -1/2$. The superconducting dome is distinguishable in Fig. 5a directly beneath this Landau

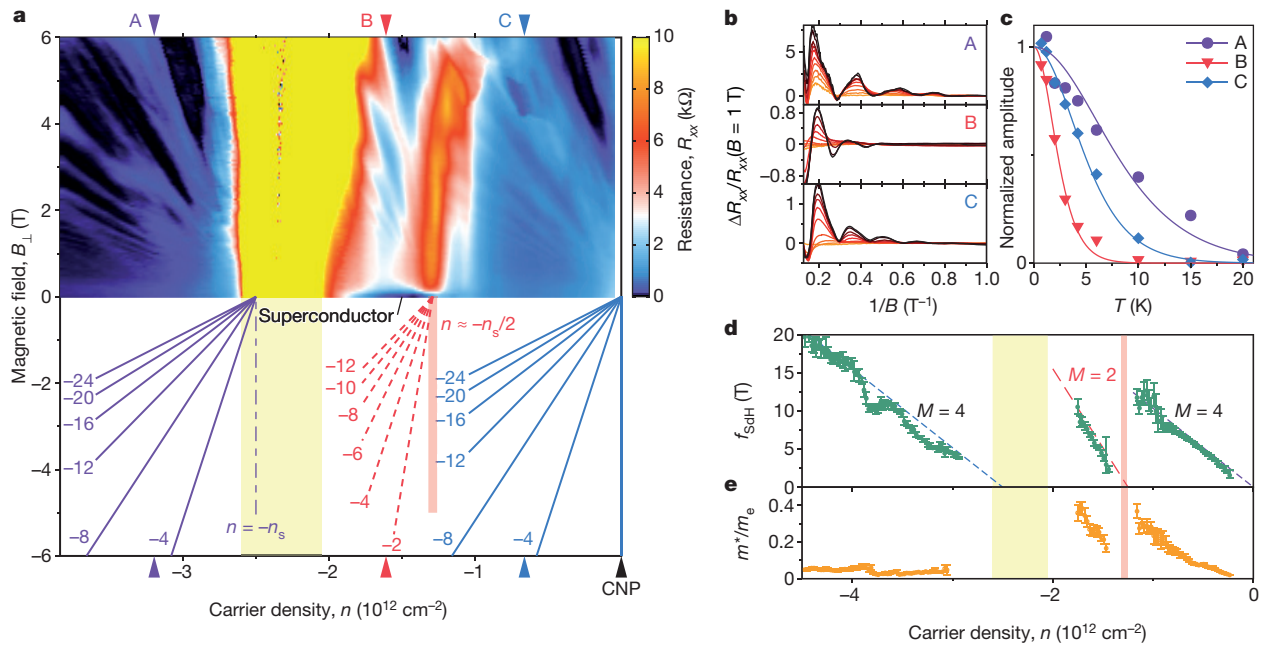


Figure 5 | Quantum oscillations in magic-angle TBG at high fields.

a, Resistance R_{xx} versus density n (hole-doped side with respect to charge neutrality) and B_{\perp} in device M2. The lower half of the diagram shows the Landau-level structure deduced from the oscillations. The blue Landau fan, which originates from the charge neutrality point (CNP), and the purple Landau fan, which originates from the superlattice density ($n = -n_s$, yellow shaded region), illustrate the filling-factor sequences $-4, -8, -12, \dots$ expected from the single-particle band structure with four-fold spin and valley degeneracies. The additional red fan, which originates from $-n_s/2$ (red shaded region), instead has a filling-factor sequence of $-2, -4, -6, \dots$

fan, being very close to zero field and next to the correlated insulating region. Unlike commonly observed broken-symmetry states that split from a single degenerate Landau level into multiple levels, the halved filling factors appear to be intrinsic to the fan, holding down to the lowest magnetic field at which oscillations are still visible. Fractional values for s have been reported in graphene superlattices as a result of Hofstadter's butterfly, which typically occurs in much stronger magnetic fields (greater than 10 T) but becomes obvious only at the intersection of Landau levels with different integer s (refs 29–31). Therefore, the physics of Hofstadter's butterfly cannot explain the additional stand-alone fan observed here, which appears at fields as low as 1 T. Furthermore, the halving of the filling factors and s is unlikely to be explained in a non-interacting picture of unit-cell doubling due to strain or to the formation of a charge density wave, in which case either spin or valley degeneracy must be broken. We observed the same Landau level sequence in two other magic-angle TBG devices, so it is robust against small variations in twist angle and consistent across samples (Methods, Extended Data Fig. 2).

To study the non-trivial origin of the Landau fan near half-filling further, we measured the effective mass from the temperature-dependent quantum oscillation amplitude according to the Lifshitz–Kosevich formula (Methods). In Fig. 5b, c we show the oscillations and oscillation amplitudes at three different densities (indicated by arrows in Fig. 5a). In Fig. 5d, e we show the oscillation frequency f_{SDH} and the effective mass extracted by fitting the oscillation amplitudes to the Lifshitz–Kosevich formula. The dependence of f_{SDH} on carrier density n provides another perspective on the oscillations because the value of $M = \phi_0 \Delta n / \Delta f_{SDH}$ extracted from the slope $\Delta n / \Delta f_{SDH}$ provides the number of degenerate Fermi pockets M directly. The experimental data clearly fit to $M = 4$ near the charge neutrality point and for densities beyond the superlattice gap, whereas $M = 2$ for the quantum oscillations that start near the correlated insulator state and right above the

superconducting dome. The effective mass of the anomalous oscillations is about $(0.2\text{--}0.4)m_e$, where m_e is the bare electron mass. This mass is much larger than the mass near charge neutrality (about $0.1m_e$) and beyond the superlattice gap (about $0.05m_e$) at the same Δn , where Δn is density relative to the value of n at which $f_{SDH} = 0$ in Fig. 5d. The quantum oscillations above the superconducting dome clearly indicate the existence of small Fermi surfaces that originate from the correlated insulating state, which have areas proportional to $n' = |n| - n_s/2$, rather than of a large Fermi surface with an area that corresponds to the density $|n|$ itself. The Hall measurements shown in Extended Data Fig. 3 also support this conclusion. Notably, similar small Fermi pockets that do not correspond to any pockets in the single-particle Fermi surface have been observed in underdoped cuprates, although their origin is debated^{32–34}. Among the possibilities, the small Fermi surface that we observe could be the Fermi surface of quasiparticles that are created by doping a Mott insulator^{6,35}. On the other hand, the halved degeneracy might be related to spin–charge separation, as predicted in a doped Mott insulator³⁵. More experimental and theoretical work is needed to clarify the origin of the quantum oscillations.

Discussion

The appearance of both superconductor and correlated insulator phases in the flat bands of magic-angle TBG at such a small carrier density cannot be explained by weak-coupling BCS theory. The carrier density that is responsible for $T_c = 1.7$ K is extremely small according to the quantum oscillation measurements, merely $n' = 1.5 \times 10^{11}$ cm $^{-2}$ at optimal doping. To place this in the context of other superconductors, in Fig. 6 we plot T_c against T_F on a logarithmic scale for various materials, where T_F is the Fermi temperature. T_F is proportional to the two-dimensional carrier density n_{2D} , which the quantum oscillations data show to be equivalent to n' for the

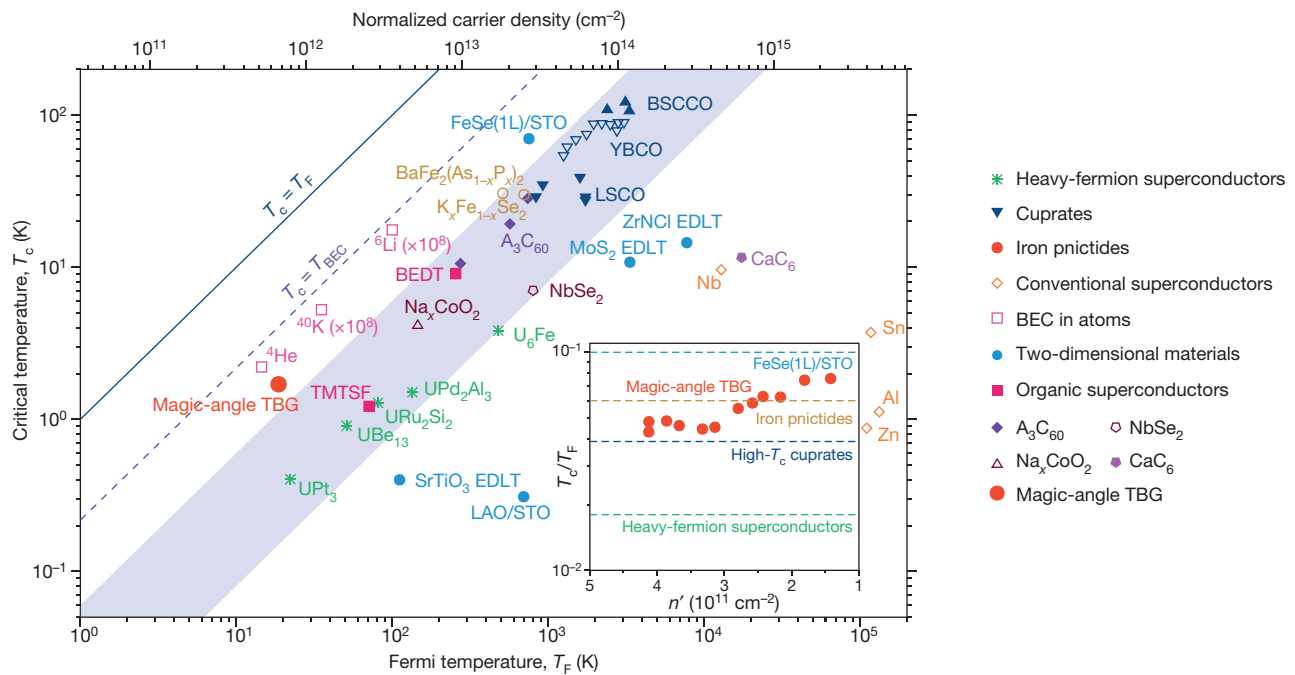


Figure 6 | Superconductivity in the strong-coupling limit. Logarithmic plot of critical temperature T_c versus Fermi temperature T_F for various superconductors³⁶. The top axis is the corresponding two-dimensional carrier density n_{2D} for two-dimensional materials or n_{3D}^2 for three-dimensional materials, normalized by the effective mass m^*/m_e and the Fermi surface degeneracy g (and a constant factor of $1/1.52$ for the three-dimensional density). Two-dimensional superconductors are represented by filled circles; other symbols represent three-dimensional (but potentially two-dimensional-like) superconductors. For comparison, we also plot $T_{BEC} = 1.04\hbar n_{3D}/m^*$ for a three-dimensional bosonic gas (dashed line). Bose–Einstein condensation temperatures in ^4He , paired fermionic ^{40}K and paired fermionic ^6Li are shown as open pink squares^{36,44} (T_c and T_F have both been multiplied by 10^8 for ^{40}K and ^6Li). The point for magic-

angle TBG (large red filled circle) is calculated from the two-dimensional density and the effective mass obtained from quantum oscillations (Fig. 5d, e) at the optimal doping ($n_{2D} = 1.5 \times 10^{11} \text{ cm}^{-2}$ and $m^* = 0.2m_e$), using $g = 1$ to account for the halved degeneracy. Data for other materials are from refs 36,45–54. The blue shaded region is the approximate region in which almost all known unconventional superconductors lie. The inset shows the variation in T_c/T_F as a function of doping n' for magic-angle TBG (red filled circles). The horizontal dashed lines are the approximate T_c/T_F values of the corresponding material. YBCO, $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$; LSCO, $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$; BSCCO, $\text{Bi}_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_y$; LAO, LaAlO_3 ; STO, SrTiO_3 ; 1L, single layer; EDLT, electric double-layer transistor; BEDT, bisethylenedithiol; TMTSE, tetramethyltetraselenafulvalene.

superconducting dome region of magic-angle TBG³⁶. Most unconventional superconductors have T_c/T_F values of about 0.01–0.05, whereas all of the conventional BCS superconductors lie on the far right in the plot, with much smaller ratios. Magic-angle TBG is located above the trend line on which most cuprates, heavy-fermion and organic superconductors lie, with a T_c/T_F value approaching that of the recently observed exotic FeSe monolayer on SrTiO_3 (Fig. 6 inset). This finding strongly suggests that the superconductivity in magic-angle TBG originates from electron correlations instead of weak electron–phonon coupling. One other frequently compared temperature is the Bose–Einstein condensation temperature for a three-dimensional boson gas T_{BEC} , assuming that all particles in the occupied Fermi sea pair up and condense. Cuprates and other unconventional superconductors typically have T_c/T_{BEC} ratios of roughly 0.1–0.2. The T_c/T_{BEC} ratio for magic-angle TBG is estimated to be up to 0.37, indicating very strong electron–electron interactions and possibly close proximity to the BCS–BEC crossover. This behaviour is in agreement with the fact that the coherence length in magic-angle TBG ($\xi \approx 50 \text{ nm}$ at optimal doping) is of the same order of magnitude as the average inter-particle distance, $(n')^{-1/2} \approx 26 \text{ nm}$.

The realization of unconventional superconductivity in a graphene superlattice establishes magic-angle TBG as a relatively simple, clean, accessible and, most importantly, highly tunable material, which could be used to study correlated electron physics. The interactions in magic-angle TBG could possibly be further fine-tuned by the twist angle and by the application of perpendicular electric fields by means of differential gating^{18,37}. Moreover, T_c could possibly be enhanced further by applying pressure to the graphene superlattice to increase the interlayer

hybridization or by coupling different magic-angle TBG structures to induce Josephson coupling in the vertical direction³⁸. Similar magic-angle superlattices and flat-band electronic structures could also be realized with other two-dimensional materials or lattices to investigate strongly correlated systems with different properties.

Finally, despite several apparent similarities between magic-angle TBG and cuprates, there are key differences between the realizations of them. First, the valley degree of freedom in the underlying graphene lattices leads to an extra degeneracy, resulting in two carriers per superlattice unit cell at half-filling in the parent correlated insulator state. Higher quality devices and fine tuning may lead to superconductivity near the regions corresponding to one and three carriers per unit cell. Second, in magic-angle TBG the underlying superlattice is triangular, which should have a fundamental influence on the type of spin-singlet ground state it can host, owing to magnetic frustration. The lattice symmetry should also impose limitations on the possible superconducting pairing symmetry in magic-angle TBG; further experiments, for example, involving tunnelling and Josephson heterojunctions, are required to confirm this³⁹. Various pairing symmetries, including $(d + id')$ -wave, $(p_x + ip_y)$ -wave and spin-triplet s -wave symmetries, have been predicted theoretically in the hypothetical superconductivity of monolayer or few-layer graphene^{40–42}. If the mechanism for superconductivity in magic-angle TBG is indeed related to the correlated half-filling insulating state, as is the case in $d_{x^2-y^2}$ -wave cuprates, then the pairing symmetry might be chiral $(d + id')$ -wave, to satisfy the underlying triangular symmetry of the superlattice. We anticipate that further experimental and theoretical work on magic-angle TBG and related magic-angle superlattices will

provide insights into the key factors that govern unconventional superconductivity, and bring us closer to realizing tunable quantum spin liquids⁴³.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 February; accepted 26 February 2018.

Published online 5 March 2018.

- Rajagopal, K. & Wilczek, F. in *At the Frontier of Particle Physics* (ed. Shifman, M.) Vol. 3, 2061–2151 (World Scientific, 2001).
- v. Löhneysen, H., Rosch, A., Vojta, M. & Wölfle, P. Fermi-liquid instabilities at magnetic quantum phase transitions. *Rev. Mod. Phys.* **79**, 1015–1075 (2007).
- Stormer, H. L. Nobel Lecture: The fractional quantum Hall effect. *Rev. Mod. Phys.* **71**, 875–889 (1999).
- Pfleiderer, C. Superconducting phases of *f*-electron compounds. *Rev. Mod. Phys.* **81**, 1551–1624 (2009).
- Ishiguro, T., Yamaji, K. & Saito, G. *Organic superconductors* 2nd edn (Springer, 1998).
- Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78**, 17–85 (2006).
- Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. From quantum matter to high-temperature superconductivity in copper oxides. *Nature* **518**, 179–186 (2015).
- Stewart, G. R. Superconductivity in iron compounds. *Rev. Mod. Phys.* **83**, 1589–1652 (2011).
- Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold gases. *Rev. Mod. Phys.* **80**, 885–964 (2008).
- Mazurenko, A. *et al.* A cold-atom Fermi–Hubbard antiferromagnet. *Nature* **545**, 462–466 (2017).
- Wang, Z., Liu, C., Liu, Y. & Wang, J. High-temperature superconductivity in one-unit-cell FeSe films. *J. Phys. Condens. Matter* **29**, 153001 (2017).
- Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
- Suárez Morell, E., Correa, J. D., Vargas, P., Pacheco, M. & Barticevic, Z. Flat bands in slightly twisted bilayer graphene: tight-binding calculations. *Phys. Rev. B* **82**, 121407 (2010).
- Moon, P. & Koshino, M. Energy spectrum and quantum Hall effect in twisted bilayer graphene. *Phys. Rev. B* **85**, 195458 (2012).
- Fang, S. & Kaxiras, E. Electronic structure theory of weakly interacting bilayers. *Phys. Rev. B* **93**, 235153 (2016).
- Trambly de Laissardiére, G., Mayou, D. & Magaud, L. Numerical studies of confined states in rotated bilayers of graphene. *Phys. Rev. B* **86**, 125413 (2012).
- Cao, Y. *et al.* Superlattice-induced insulating states and valley-protected orbits in twisted bilayer graphene. *Phys. Rev. Lett.* **117**, 116804 (2016).
- Cao, Y. *et al.* Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, <https://doi.org/10.1038/nature26154> (2018).
- Lopes dos Santos, J. M. B., Peres, N. M. R. & Castro Neto, A. H. Continuum model of the twisted graphene bilayer. *Phys. Rev. B* **86**, 155449 (2012).
- Kim, K. *et al.* van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
- Kim, K. *et al.* Tunable moiré bands and strong correlations in small-twist-angle bilayer graphene. *Proc. Natl Acad. Sci. USA* **114**, 3364–3369 (2017).
- Wang, L. *et al.* One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).
- Tinkham, M. *Introduction to Superconductivity* (Courier Corporation, 1996).
- Saito, Y., Nojima, T. & Iwasa, Y. Highly crystalline 2D superconductors. *Nat. Rev. Mater.* **2**, 16094 (2016).
- Mott, N. F. *Metal-Insulator Transitions* (Taylor and Francis, 1990).
- Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70**, 1039–1263 (1998).
- Klemm, R. A. & Luther, A. Theory of the upper critical field in layered superconductors. *Phys. Rev. B* **12**, 877–891 (1975).
- Goldman, A. M. in *BCS: 50 Years* (eds Cooper, L. N. & Feldman, D.) 255–275 (World Scientific, 2011).
- Hunt, B. *et al.* Massive Dirac fermions and Hofstadter butterfly in a van der Waals heterostructure. *Science* **340**, 1427–1430 (2013).
- Ponomarenko, L. A. *et al.* Cloning of Dirac fermions in graphene superlattices. *Nature* **497**, 594–597 (2013).
- Dean, C. R. *et al.* Hofstadter’s butterfly and the fractal quantum Hall effect in moiré superlattices. *Nature* **497**, 598–602 (2013).
- Yelland, E. A. *et al.* Quantum oscillations in the underdoped cuprate YBa₂Cu₄O₈. *Phys. Rev. Lett.* **100**, 047003 (2008).
- Bangura, A. F. *et al.* Small Fermi surface pockets in underdoped high temperature superconductors: observation of Shubnikov–de Haas oscillations in YBa₂Cu₄O₈. *Phys. Rev. Lett.* **100**, 047004 (2008).
- Jaudet, C. *et al.* de Haas–van Alphen oscillations in the underdoped high-temperature superconductor YBa₂Cu₃O_{6.5}. *Phys. Rev. Lett.* **100**, 187005 (2008).
- Kaul, R. K., Kim, Y. B., Sachdev, S. & Senthil, T. Algebraic charge liquids. *Nat. Phys.* **4**, 28–31 (2008).
- Uemura, Y. J. Condensation, excitation, pairing, and superfluid density in high-*T_c* superconductors: the magnetic resonance mode as a roton analogue and a possible spin-mediated pairing. *J. Phys. Condens. Matter* **16**, S4515–S4540 (2004).
- Gonzalez-Arraga, L. A., Lado, J. L., Guinea, F. & San-Jose, P. Electrically controllable magnetism in twisted bilayer graphene. *Phys. Rev. Lett.* **119**, 107201 (2017).
- Yankowitz, M. *et al.* Dynamic band-structure tuning of graphene moiré superlattices with pressure. *Nature* (in the press); preprint at <https://arxiv.org/abs/1707.09054> (2017).
- Tsuei, C. C. & Kirtley, J. R. Pairing symmetry in cuprate superconductors. *Rev. Mod. Phys.* **72**, 969 (2000).
- Nandkishore, R., Levitov, L. S. & Chubukov, A. V. Chiral superconductivity from repulsive interactions in doped graphene. *Nat. Phys.* **8**, 158–163 (2012).
- Uchoa, B. & Castro Neto, A. H. Superconducting states of pure and doped graphene. *Phys. Rev. Lett.* **98**, 146801 (2007).
- Hosseini, M. V. & Zareyan, M. Unconventional superconducting states of interlayer pairing in bilayer and trilayer graphene. *Phys. Rev. B* **86**, 214503 (2012).
- Balents, L. Spin liquids in frustrated magnets. *Nature* **464**, 199–208 (2010).
- Ku, M. J. H., Sommer, A. T., Cheuk, L. W. & Zwierlein, M. W. Revealing the superfluid lambda transition in the universal thermodynamics of a unitary fermi gas. *Science* **335**, 563–567 (2012).
- Qian, T. *et al.* Absence of a Holeylike Fermi surface for the iron-based K_{0.8}Fe_{1.7}Se₂ superconductor revealed by angle-resolved photoemission spectroscopy. *Phys. Rev. Lett.* **106**, 187001 (2011).
- Hashimoto, T. *et al.* Sharp peak of the zero-temperature penetration depth at optimal composition in BaFe₂(As_{1-x}P_x)₂. *Science* **336**, 1554–1557 (2012).
- Saito, Y., Kasahara, Y., Ye, J., Iwasa, Y. & Nojima, T. Metallic ground state in an ion-gated two-dimensional superconductor. *Science* **350**, 409–413 (2015).
- Ye, J. T. *et al.* Superconducting dome in a gate-tuned band insulator. *Science* **338**, 1193–1196 (2012).
- Peelaers, H. & Van de Walle, C. G. Effects of strain on band structure and effective masses in MoS₂. *Phys. Rev. B* **86**, 241401(R) (2012).
- Caviglia, A. D. *et al.* Electric field control of the LaAlO₃/SrTiO₃ interface ground state. *Nature* **456**, 624–627 (2008).
- McCollam, A. *et al.* Quantum oscillations and subband properties of the two-dimensional electron gas at the LaAlO₃/SrTiO₃ interface. *APL Mater.* **2**, 022102 (2014).
- Ueno, K. *et al.* Electric-field-induced superconductivity in an insulator. *Nat. Mater.* **7**, 855–858 (2008).
- Weller, T. E., Ellerby, M., Saxena, S. S., Smith, R. P. & Skipper, N. T. Superconductivity in the intercalated graphite compounds C₆Yb and C₆Ca. *Nat. Phys.* **1**, 39–41 (2005).
- Valla, T. *et al.* Anisotropic electron-phonon coupling and dynamical nesting on the graphene sheets in superconducting CaC₆ using angle-resolved photoemission spectroscopy. *Phys. Rev. Lett.* **102**, 107007 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge discussions with R. Ashoori, S. Carr, R. Comin, L. Fu, P. A. Lee, L. Levitov, K. Rajagopal, S. Todadri, A. Vishwanath and M. Zwierlein. This work was primarily supported by the Gordon and Betty Moore Foundation’s EPIQS Initiative through grant GBMF4541 and the STC Center for Integrated Quantum Materials (NSF grant number DMR-1231319) for device fabrication, transport measurements and data analysis (Y.C., P.J.-H.), and theoretical calculations (S.F.). Data analysis by V.F. was supported by AFOSR grant number FA9550-16-1-0382. K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by MEXT, Japan and JSPS KAKENHI grant numbers JP15K21722 and JP25106006. This work made use of the Materials Research Science and Engineering Center Shared Experimental Facilities, supported by the NSF (DMR-0819762), and of Harvard’s Center for Nanoscale Systems, supported by the NSF (ECS-0335765). E.K. acknowledges additional support by ARO MURI award W911NF-14-0247.

Author Contributions Y.C. fabricated samples and performed transport measurements. Y.C., V.F. and P.J.-H. performed data analysis and discussed the results. P.J.-H. supervised the project. S.F. and E.K. provided numerical calculations. K.W. and T.T. provided hexagonal boron nitride samples. Y.C., V.F. and P.J.-H. co-wrote the manuscript with input from all co-authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.J.-H. (pij@mit.edu) or Y.C. (caoyuan@mit.edu).

Reviewer Information *Nature* thanks E. Mele, J. Robinson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Sample preparation. The devices were fabricated using a modified dry-transfer technique^{17,18,20}. Monolayer graphene and hexagonal boron nitride (about 10–30 nm thick) were exfoliated on SiO₂/Si chips and high-quality flakes were picked using optical microscopy and atomic force microscopy. We used a poly(bisphenol A carbonate) (PC)/polydimethylsiloxane (PDMS) stack on a glass slide mounted on a custom-made micro-positioning stage to pick up a hexagonal boron nitride flake at 90 °C, and then used the van der Waals force between hexagonal boron nitride and graphene to tear a graphene flake at room temperature. The separated graphene pieces were rotated manually by a twist angle of about 1.2°–1.3° and stacked together again, which resulted in a controlled TBG structure. The stack was encapsulated with another hexagonal boron nitride flake on the bottom and released onto a metal gate at 160 °C. We did not perform any heat annealing after this step because we found that TBG tended to relax to Bernal-stacked bilayer graphene at high temperatures. The final device geometry was defined by using electron-beam lithography and reactive ion etching. Electrical connections were made to the TBG by Cr/Au edge-contacted leads²².

Measurements. Transport measurements were performed in a dilution refrigerator with a base temperature of 70 mK, except for the temperature-dependent quantum oscillations, which were measured in a ³He fridge.

We used standard low-frequency lock-in techniques with an excitation frequency of about 5–10 Hz and an excitation current of about 0.4–5 nA. The current flowing through the sample was amplified by a current pre-amplifier and measured by the lock-in amplifier. The four-probe voltage was amplified by a voltage pre-amplifier at $\times 1,000$ gain and measured by another lock-in amplifier.

The twist angle of the devices was determined from the transport measurements at low temperatures¹⁸. In brief, a rough estimate of the twist angle is provided by the carrier density of the superlattice gaps at $\pm n_s$, which present as strongly insulating states. To refine this estimate, the Landau levels that appear at high magnetic fields were fitted to the Wannier diagram, which gives the twist angle with an uncertainty of about 0.01°–0.02°.

Extracting the quantum oscillation frequency and effective mass. The effective mass in device M2 was extracted using the standard Lifshitz–Kosevich formula, which relates the temperature-dependence of resistance change $\Delta R_{xx}(T)$ to the cyclotron mass m^* (at a given magnetic field B_{\perp}):

$$\Delta R_{xx}(T) \propto \frac{\chi}{\sinh(\chi)}, \quad \chi = \frac{2\pi^2 k_B T m^*}{\hbar e B_{\perp}}$$

For each gate voltage (carrier density), we measured the R_{xx} – B_{\perp} curves at different temperatures, normalized them by their low-field values and subtracted a common polynomial background in B_{\perp} . Examples of the curves are shown in Fig. 5b. The oscillation frequencies shown in Fig. 5d were extracted from these curves plotted versus $1/B_{\perp}$. From the temperature-dependent amplitude of the most prominent peak, we extracted m^* using the above equation (Fig. 5e). The error bars in Fig. 5d, e represent 90% confidence intervals of the fit.

Commensuration and twist angle. Mathematically, in a twisted moiré system, the lattice is strictly periodic only when the twist angle satisfies a specific relation such that lattice registration order is perfectly recovered in a finite distance. These special cases are termed ‘commensurate’ structures. One important parameter in commensurate TBG structures is r , which can be intuitively understood as the number of ‘apparent’ moiré pattern wavelengths that it takes to recover the lattice periodicity fully^{19,55}. The simplest commensurate structures with $r=1$ are called ‘minimal’ structures. These structures have exactly one moiré spot per unit cell. In TBG, as well as the minimal structures, which occur only at discrete angles, there are other commensurate structures that are arbitrarily close to any given angle θ with large r . However, at small twist angles, the evolution of the band structure of TBG can be viewed as semi-continuous; that is, an infinitesimal change in twist angle does not have a substantial effect on the band structure even though the lattice could be in a different family of commensurate structures (different r)¹⁹. In other words, the TBG system can be well approximated by a continuum model, as originally proposed in ref. 12, and the physics in minimal structures is representative of all nearby commensurate structures¹². In our experiments, we do not expect the lattice to be in perfect commensuration, owing to disorder and intrinsic randomness due to the fabrication process. However, we think that the continuum model can faithfully represent the realistic TBG system in which any commensuration effect has been smoothed out.

We deduced the size of the moiré unit cell and the twist angle on the basis of the density of the superlattice gaps $\pm n_s$ (± 4 electrons per moiré unit cell), and then cross-checked the twist angle with the Landau levels observed at high magnetic fields. $\pm n_s$ are the only multiples of n_c that correspond to Fermi energies located within single-particle band gaps and therefore exhibit strong insulating behaviour.

For twist angles above about 0.9°–1°, the band structure at energies higher than these gaps is strongly overlapping and no single particle gaps at $\pm 2n_s, \pm 3n_s, \dots$ appear^{12–14,19,56}. The experimentally measured values for the single-particle insulating gaps that we observe are in the approximately 30–60-meV range^{17,18}. However, below about 0.9°–1°, the superlattice gaps at $\pm n_s$ close and there is no single-particle gap at any energy in the system^{21,56}. In this regime, there are Dirac-like bands that cross at $\pm 2n_s$ which might be responsible for the resistance peaks observed in devices with very small twist angles, although possible interaction effects may enhance these peaks²¹. The states observed in very-low-twist devices are clearly different from the strong insulating gaps observed here and previously¹⁸. There is a marked change in the band structure at about 0.9°–1° (depending on the parameters of the model being used), which leads to a transition from single-particle gaps at $\pm n_s$ to resistive states at $\pm 2n_s$. This crossover can be observed clearly in Supplementary Video, in which we show an evolution of the band structure of TBG from $\theta=3^\circ$ to $\theta=0.8^\circ$. The data in the video were calculated using the continuum model¹².

Possible effects due to finite electrical fields. It has been shown that by applying a perpendicular electrical field to Bernal-stacked bilayer graphene, topological states can emerge on the AB/BA stacking boundaries while the bulk of the AB and BA regions remains gapped^{57–59}. In small-angle TBG, a similar effect can alter the band structure because the AA-stacked regions in the moiré pattern are interconnected by the AB/BA stacking boundaries. This effect has been observed recently in scanning tunnelling experiments on ultrasmall-twist-angle samples⁶⁰.

The question then arises of how the flat bands in magic-angle TBG are affected by the network of topological boundaries when a residual electrical field is present. Theoretical work on $\theta=1.5^\circ$ TBG has shown that when an inter-layer potential difference of $\Delta V=300$ mV is applied the low-energy superlattice bands become even flatter and the electronic states become more localized³⁷. Therefore, there is good reason to believe that the flat-band physics presented here holds even when a perpendicular electric field is present, because the electric field will probably render the band structure even more localized and correlated as the twist angle approaches the magic angle. In our experiments, we estimate that the potential difference between the two layers induced by our gate voltage is at most about 50 mV, and probably much less, owing to screening. Any possible effects of the residual electric field should be minimal.

Phase-coherent transport behaviour in superconducting magic-angle TBG. In Fig. 3a, b we observe oscillatory behaviour in the measured longitudinal resistance R_{xx} as a function of perpendicular magnetic field B_{\perp} when the charge density is close to the boundary between the half-filling insulating state and the superconducting states. The oscillations are most clearly seen for $n \approx -1.70 \times 10^{12} \text{ cm}^{-2}$ to $n \approx -1.60 \times 10^{12} \text{ cm}^{-2}$ and $n \approx -1.50 \times 10^{12} \text{ cm}^{-2}$ to $n \approx -1.47 \times 10^{12} \text{ cm}^{-2}$ in device M1.

In Extended Data Fig. 1a, b we show the differential resistance dV_{xx}/dI versus bias current I and perpendicular magnetic field B_{\perp} . At zero bias current, the oscillations of the differential resistance with B_{\perp} shown correspond to line cuts in Fig. 3a at densities of $n \approx -1.48 \times 10^{12} \text{ cm}^{-2}$ (Extended Data Fig. 1a) and $n \approx -1.68 \times 10^{12} \text{ cm}^{-2}$ (Extended Data Fig. 1b). The critical current, above which the superconductor becomes normal, oscillates with B_{\perp} at the same frequency, as can be visualized by the bright peaks in Extended Data Fig. 1a, b. The oscillation period is $\Delta B=22.5$ mT in Extended Data Fig. 1a and about $\Delta B=4$ mT in Extended Data Fig. 1b.

The fact that the critical current is maximum at zero B_{\perp} and oscillates at periodic intervals of the magnetic field suggests the existence of Josephson junction arrays—in the simplest case, a superconducting quantum interference device (SQUID)-like superconducting loop, around a normal or insulating island²³. It is unclear whether this inhomogeneous behaviour is a result of sample disorder or a coexistence of two different phases (such as the superconducting phase and the correlated insulator phase). Owing to the two-dimensional nature of our devices, the detailed current distribution in the device cannot be uniquely determined at this moment by transport measurements; however, from the oscillation period we deduce the effective loop area of the SQUID approximately using $S=\Phi_0/\Delta B$, where $\Phi_0=h/(2e)$ is the superconducting quantum flux. (Note the difference between $\Phi_0=h/e$ for the quantum Hall effect and $\Phi_0=h/(2e)$ for superconductivity.) For the experimental data in Extended Data Fig. 1a, b, we obtain areas of $S=0.09 \mu\text{m}^2$ and $S=0.5 \mu\text{m}^2$, respectively. By comparison, the total device area between the voltage probes is approximately $1 \mu\text{m}^2$.

Using a simple model of a SQUID with a phenomenological decay of the oscillation amplitude at higher magnetic fields, we attempt to reproduce the observed oscillations qualitatively using numerical simulations. In Extended Data Fig. 1c we show the simulated I – B_{\perp} map of the differential resistance for a SQUID with area $S=0.09 \mu\text{m}^2$, with the same critical current $I_{c1}=I_{c2}=7$ nA in the two branches, corresponding to the experimental data in Extended Data Fig. 1a. In Extended Data

Fig. 1d we show the simulation for an asymmetric SQUID with area $S = 0.5 \mu\text{m}^2$ and critical currents of $I_{c1} = 6 \text{ nA}$ and $I_{c2} = 10 \text{ nA}$ for the two branches, which account for the partial cancellation of the critical current at low fields (that is, the total critical current does not reach zero in an oscillation) seen in Extended Data Fig. 1b. These simulations provide a qualitative perspective on the oscillatory phenomenon; the actual supercurrent distribution is probably much more complex and will need to be established via magnetic imaging techniques. However, our data indicate that the superconducting behaviour that we observe is indeed a phase-coherent phenomenon. Although we did not fabricate SQUID devices deliberately using magic-angle TBG, these periodic oscillations of the critical current in B_{\perp} are probably a result of the Josephson effect through a superconductor with insulating puddles, further confirming the existence of superconductivity in magic-angle TBG.

Induced superconductivity in graphene and graphene-based systems through proximity to another superconductor has been demonstrated, and graphene-based Josephson junctions continue to be explored^{61–63}. Superconductivity in graphene induced by proximity to a high- T_c superconductor has been reported recently, and indications of induced unconventional pairing have been observed^{64,65}.

Supplementary quantum oscillation data and low-field Hall effect. In Extended Data Fig. 2 we show magneto-transport data for device M1 and another magic-angle device D1. Both devices show evidence for the existence of an extra Landau fan with a degeneracy of $M = 2$ that emerges from the half-filling insulating states. All of the magic-angle devices that we have measured so far display quantum oscillations that correspond to emergent quasiparticles on one side of the half-filling states—the one that is away from the charge neutrality point (that is, $n < -n_s/2$ for $E_F < 0$ and $n > n_s/2$ for $E_F > 0$; see ref. 18 for the $E_F > 0$ data)—but not the other ($n > -n_s/2$ or $n < n_s/2$). The Hall measurements reported below exhibit a similar asymmetry around the half-filling state. This universally asymmetric behaviour, regardless of the twist angle, might be explained if the effective mass of the quasiparticles on the side closer to charge neutrality is much larger, and therefore the corresponding quasiparticle has a much lower mobility, so that the quantum oscillations cannot be observed and their contribution to the Hall effect becomes negligible. Further theoretical work could potentially shed more light on the true nature of the many-body energy gap and the related quasiparticles.

We determined the Fermi surface area in the magic-angle TBG devices using the Shubnikov–de Hass oscillation frequency in a magnetic field (Fig. 5). We find that oscillations emerge from the correlated insulating state at half-filling $n = -n_s/2$, and the oscillation frequency indicates small Fermi pockets associated with a shifted density of $n' = |n| - n_s/2$.

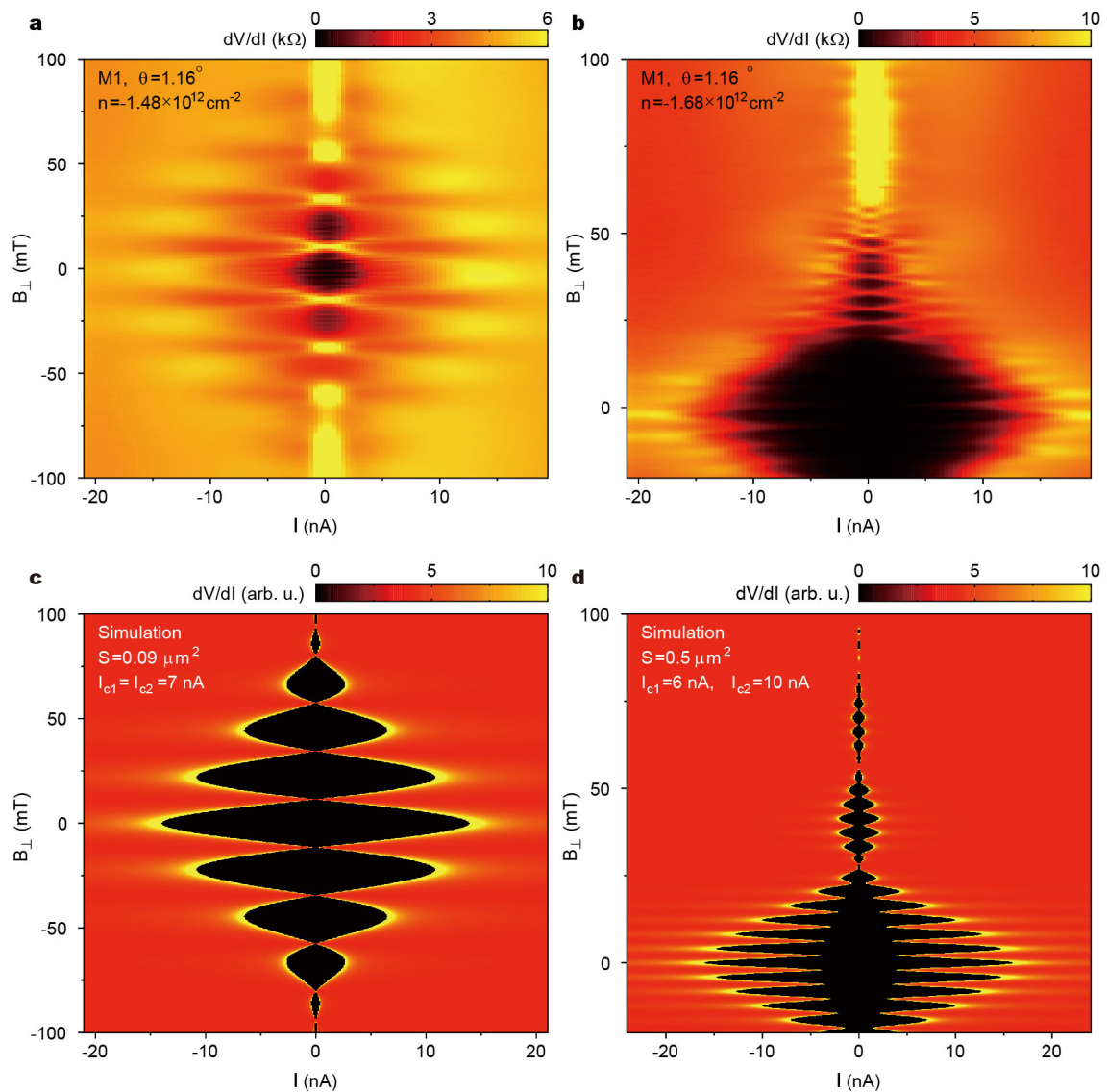
In Extended Data Fig. 3 we show another measurement of the transport carrier density via the low-field Hall effect measured up to $\pm 1 \text{ T}$. The measured Hall density, given by $n_H = -(1/e)(dR_{xy}/dB_{\perp})_{B_{\perp}=0}^{-1}$, provides an independent measurement of the carrier density in the system. In both devices, at a temperature

of 0.4 K we observe that, whereas the Hall density follows the gate-induced density closely ($n_H = n$) near charge neutrality and up to the half-filling insulating states at $|n| = n_s/2$, it ‘resets’ to a much smaller value beyond $|n| = n_s/2$. The Hall density beyond these points behaves as if the charge carriers that contribute to transport are just those added beyond $|n| = n_s/2$, and roughly follows $n_H = n + n_s/2$ for $n < -n_s/2$ and $n_H = n - n_s/2$ for $n > n_s/2$. This behaviour is in agreement with the measurements of the quantum oscillation frequency shown in Fig. 5d.

This resetting effect is quickly suppressed by raising the temperature to about 10 K. Beyond this temperature the Hall density increases monotonically towards the band edge. At these higher temperatures, the Hall density in the flat bands no longer follows $n_H = n$. This could possibly be explained by the thermal energy kT being close to the bandwidth of the flat bands, in which case the Hall coefficient must take into consideration the contributions from carriers that are thermally excited into the higher-energy, highly dispersive bands, which have opposite polarity. By contrast, up to 30 K, the Hall density measured at very high densities ($|n| > n_s$) exhibits very linear behaviour according to $|n_H| = |n| - n_s$ regardless of the temperature, which is consistent with the highly-dispersive, low-mass bands above and below the flat bands, as seen in Fig. 1c.

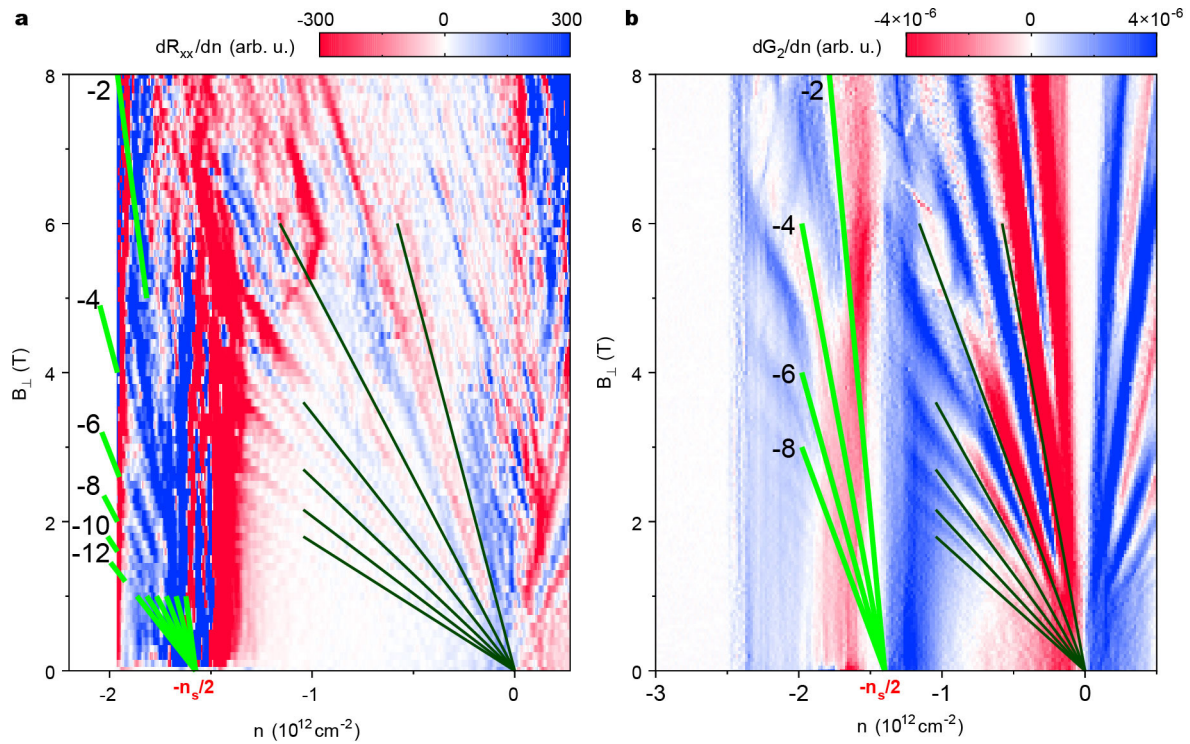
Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request.

55. Shallcross, S., Sharma, S., Kandelaki, E. & Pankratov, O. A. Electronic structure of turbostratic graphene. *Phys. Rev. B* **81**, 165105 (2010).
56. Nam, N. N. T. & Koshino, M. Lattice relaxation and energy band modulation in twisted bilayer graphene. *Phys. Rev. B* **96**, 075311 (2017).
57. Zhang, F., MacDonald, A. H. & Mele, E. J. Valley Chern numbers and boundary modes in gapped bilayer graphene. *Proc. Natl Acad. Sci. USA* **110**, 10546–10551 (2013).
58. Vaezi, A., Liang, Y., Ngai, D. H., Yang, L. & Kim, E.-A. Topological edge states at a tilt boundary in gated multilayer graphene. *Phys. Rev. X* **3**, 021018 (2013).
59. Ju, L. *et al.* Topological valley transport at bilayer graphene domain walls. *Nature* **520**, 650–655 (2015).
60. Huang, S. *et al.* Emergence of topologically protected helical states in minimally twisted bilayer graphene. Preprint at <https://arxiv.org/abs/1802.02999> (2018).
61. Heersche, H. B. *et al.* Bipolar supercurrent in graphene. *Nature* **446**, 56–59 (2007).
62. Calado, V. E. *et al.* Ballistic Josephson junctions in edge-contacted graphene. *Nat. Nanotechnol.* **10**, 761–764 (2015).
63. Bretheau, L. *et al.* Tunneling spectroscopy of Andreev states in graphene. *Nat. Phys.* **13**, 756–760 (2017).
64. Di Bernardo, A. *et al.* *p*-wave triggered superconductivity in single-layer graphene on an electron-doped oxide superconductor. *Nat. Commun.* **8**, 14024 (2017).
65. Perconte, D. *et al.* Tunable Klein-like tunnelling of high-temperature superconducting pairs into graphene. *Nat. Phys.* **14**, 25–29 (2018).



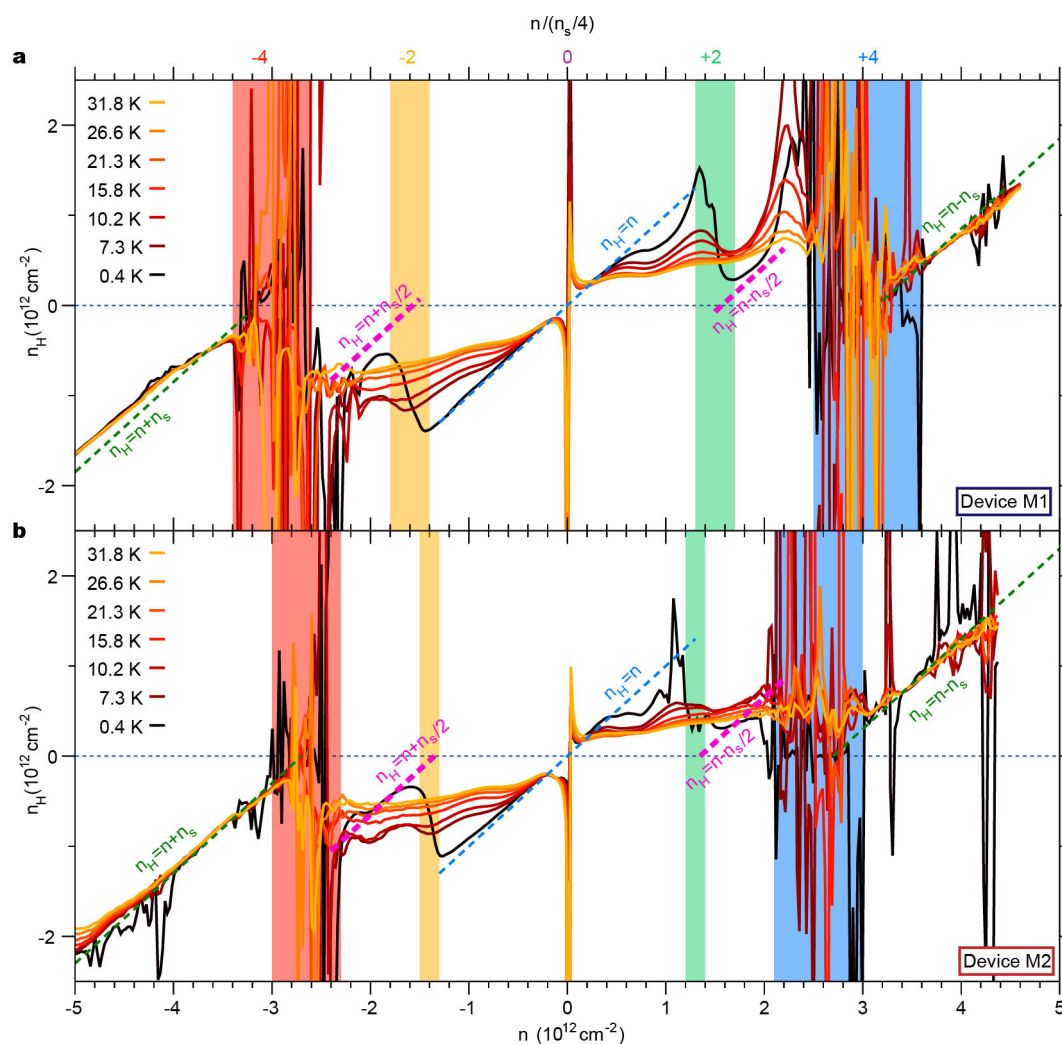
Extended Data Figure 1 | Evidence of phase-coherent transport in superconducting magic-angle TBG. **a, b,** Differential resistance dV/dI versus bias current I and perpendicular field B_\perp , at two different charge densities n , corresponding to those in Fig. 3a. Periodic oscillations are

observed in the critical current (identified approximately as the position of the bright peaks in dV/dI). **c, d,** Simulations intended to reproduce qualitatively the behaviour observed in **a** and **b**.



Extended Data Figure 2 | Supplementary quantum oscillation data.
a, b, Quantum oscillations in device M1 (**a**; $\theta = 1.16^\circ$, data shown for R_{xx}) and device D1 (**b**; $\theta = 1.08^\circ$, data shown for the two-probe conductance G_2). The first derivative with respect to the gate-defined charge density n has been taken in both cases to enhance the colour contrast. Both devices

exhibit a Landau fan that emerges from the half-filling state $-n_s/2$ and have a Landau level sequence of $-2, -4, -6, -8, \dots$, consistent with the results shown in Fig. 5. By comparison, the Landau fans that start from charge neutrality have a sequence of $-4, -8, -12, \dots$



Extended Data Figure 3 | Low-field Hall effect in magic-angle TBG.

a, b, Low-field Hall effect for devices M1 (**a**) and M2 (**b**). The Hall density $n_H = -(1/e)(dR_{xy}/dB_z)_{B_z=0}^{-1}$ is plotted as a function of the total charge density induced by the gate (n), measured at temperatures from 0.4 K to

31.8 K. Coloured vertical bars correspond to densities of $-n_s$, $-n_s/2$, $n_s/2$ and n_s for the two samples. Dashed lines are the expected Hall density if the offset given in the corresponding formula is considered.

The logic of single-cell projections from visual cortex

Yunyun Han^{1,2,3*}, Justus M. Kebschull^{4,5*}, Robert A. A. Campbell^{3*}, Devon Cowan³, Fabia Imhof³, Anthony M. Zador^{5§} & Thomas D. Mrsic-Flogel^{3,6§}

Neocortical areas communicate through extensive axonal projections, but the logic of information transfer remains poorly understood, because the projections of individual neurons have not been systematically characterized. It is not known whether individual neurons send projections only to single cortical areas or distribute signals across multiple targets. Here we determine the projection patterns of 591 individual neurons in the mouse primary visual cortex using whole-brain fluorescence-based axonal tracing and high-throughput DNA sequencing of genetically barcoded neurons (MAPseq). Projections were highly diverse and divergent, collectively targeting at least 18 cortical and subcortical areas. Most neurons targeted multiple cortical areas, often in non-random combinations, suggesting that sub-classes of intracortical projection neurons exist. Our results indicate that the dominant mode of intracortical information transfer is not based on ‘one neuron–one target area’ mapping. Instead, signals carried by individual cortical neurons are shared across subsets of target areas, and thus concurrently contribute to multiple functional pathways.

While the inputs received by a neuron drive its activity, its axonal projections determine its effects on other neurons. The axons of excitatory projection neurons that reside in layers 2 and 3 (hereafter 2/3), and 5 and 6 of the neocortex are the main conduit by which signals are exchanged between cortical areas¹. To date, no study has, to our knowledge, systematically investigated the principles by which individual neurons in any region of the mammalian neocortex distribute information to their targets. This knowledge is fundamental for deducing the logic of the communication between areas, for constraining hypotheses about neural function and for the identification of putative sub-classes of neurons. Anatomical studies in macaques, cats and mice, which are mostly based on retrograde tracing methods, indicate that there is an abundance of intracortical projection neurons in the sensory neocortex, which have axons that appear to innervate single target areas^{2–6}, raising the possibility that information may be distributed through ensembles of dedicated pathways that are functionally tailored to each target^{6–12}. For example, neurons in the mouse primary visual cortex (V1) that innervate the posteromedial (PM) or anterolateral (AL) area appear to match the spatial and temporal frequency preference of these target areas^{7,13,14}. Similarly, neurons in the mouse primary somatosensory cortex that project to either the primary motor cortex or the secondary somatosensory area comprise largely non-overlapping populations with distinct physiological and functional properties^{6,9,10}. These findings indicate that dedicated lines—specialized subpopulations of neurons that preferentially target a single downstream area (Fig. 1a, left)—may represent a fundamental mode of cortico-cortical communication. Alternatively, intracortically projection neurons could broadcast to multiple targets^{4,5,15–19}, either randomly (Fig. 1a, middle) or by targeting specific sets of areas (Fig. 1a, right). These three models of cortical architecture have different implications for communication between areas underlying sensory processing in hierarchical networks. To distinguish between these models, we used two anterograde

anatomical approaches, whole-brain fluorescence-based axonal tracing and high-throughput DNA sequencing of genetically barcoded neurons (MAPseq), to map the long-range axonal projection patterns of individual neurons in the mouse primary visual cortex, an area that distributes visual information to multiple cortical and subcortical targets^{20–22}.

Fluorescence-based tracing of single neurons

We first traced the projections of single neurons using whole-brain fluorescence-based axonal reconstructions. We used single-cell electroporation of a GFP-encoding plasmid to label up to six layer-2/3 cells in the right visual cortex of each mouse. After allowing 3–10 days for GFP expression, we imaged the axonal projections of the labelled neurons by whole-brain serial two-photon tomography with $1 \times 1 \times 10\text{-}\mu\text{m}^3$ resolution^{23,24} (Fig. 1b). We then traced each fluorescently labelled cell ($n = 71$; Fig. 1c, d) and registered each brain to the Allen Reference Atlas²⁵ (Fig. 1e, f). To assess axonal labelling with GFP, we electroporated neurons labelled retrogradely from the ipsilateral striatum, and in all cases observed axonal terminations therein ($n = 9/9$ cells; Extended Data Fig. 1), indicating a low false-negative rate of filling axon collaterals in distal targets of V1 neurons. Nonetheless, to minimize any possible contribution of incomplete axonal filling, we excluded those reconstructed V1 neurons with axon collaterals beyond V1 that terminated abruptly without branching ($n = 28$; Extended Data Fig. 2 and Supplementary Note 1), although the results below are robust to the inclusion of these cells (Extended Data Fig. 2e). We did not exclude neurons with abrupt terminations of contralaterally projecting branches (see also ref. 6), instead restricting our analysis to ipsilaterally projecting axons.

We analysed the ipsilateral projection patterns of 38 pyramidal neurons in layer 2/3, including 31 neurons in area V1 (Fig. 1g and Extended Data Figs 3, 4) and 7 neurons in nearby higher visual areas (Extended Data Fig. 5). Inspection of individual axonal arborizations of V1

¹Department of Neurobiology, School of Basic Medicine and Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. ²Institute for Brain Research, Collaborative Innovation Center for Brain Science, Huazhong University of Science and Technology, Wuhan, China. ³Biozentrum, University of Basel, 4056 Basel, Switzerland. ⁴Watson School of Biological Sciences, Cold Spring Harbor, New York, USA. ⁵Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. ⁶Sainsbury Wellcome Centre, University College London, London, UK.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

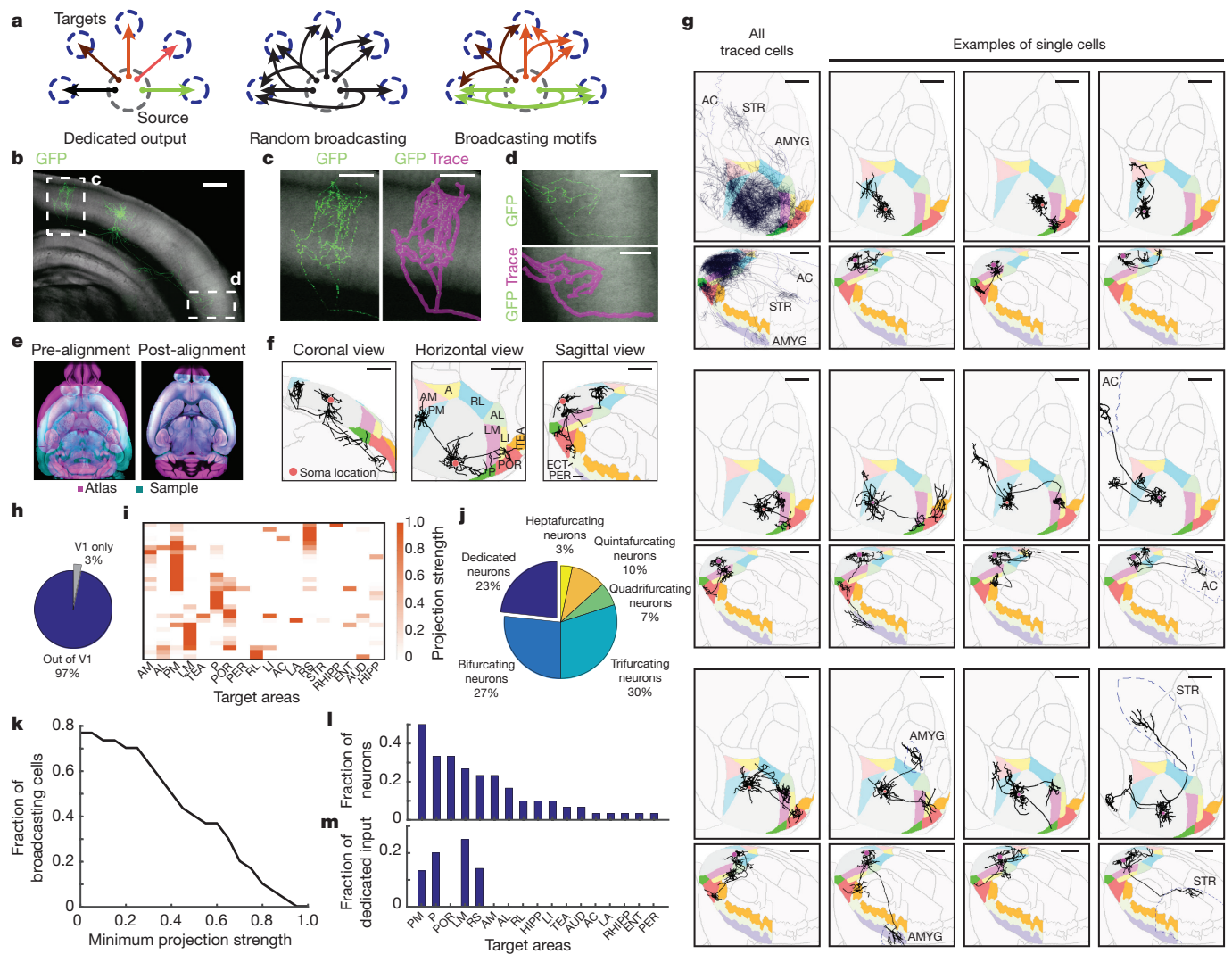


Figure 1 | Brain-wide single-cell tracing reveals the diversity of axonal projection patterns of layer-2/3 V1 neurons, with most cells projecting to more than one target area. **a**, Three hypothetical modes of information transfer from one source area to multiple target areas. Neurons (arrows) could each project to a single area (left) or to several areas either randomly (middle) or in predefined projection patterns (right). **b**, Maximum projection of the coronal view of a representative GFP-filled neuron acquired by serial-section two-photon microscopy. Auto-fluorescence from the red channel is used to show the brain's ultrastructure (grey background). Scale bar, 600 μm . $n = 71$ neurons. **c**, **d**, Higher magnification of the medial (**c**) and lateral (**d**) axonal arborization of the example neuron. Scale bars, 300 μm . **e**, Horizontal section through a sample brain (cyan) and the Allen Reference Atlas (purple) before (left) and after (right) rigid and non-rigid transformation of the brain to the atlas. **f**, Coronal, horizontal and sagittal projections of the traced example neuron overlaid in the Allen Reference Atlas space. Target cortical areas are coloured as indicated. A, anterior; AL, anterolateral; AM, anteromedial; ECT, ectothalamic; LI, laterointermediate; LM, lateromedial; P, posterior; PER, perirhinal; PM, posteromedial; POR, postrostral; RL, rostro-lateral; TEA, temporal association. Scale bars, 1 mm. **g**, Overlay of all traced single neurons (top left) and 11 example cells in Allen Reference Atlas space; horizontal view (top) and sagittal

view (bottom) of each cell. Dashed outlines label non-visual target areas. AC, anterior cingulate cortex; AMYG, amygdala; STR, striatum. Note that these images are for illustration purposes only, because a two-dimensional projection cannot faithfully capture the true three-dimensional axonal arborization pattern. Scale bar, 1 mm. **h**, The fraction of traced single neurons that project to at least one target area outside V1 is shown in blue. At least 1 mm of axonal innervation is required for an area to be considered a target. **i**, Projection pattern of all GFP-filled V1 neurons targeted randomly ($n = 31$ neurons). The colour code reflects the projection strengths of each neuron, determined as axon length per target area, normalized to the axon length in the target area receiving the densest innervation. Only brain areas that receive input from at least one neuron, as well as the striatum, are shown. AUD, auditory cortex; ENT, entorhinal; HIPP, hippocampus; LA, lateral amygdala; RHIPP, retrohippocampal region; RS, retrosplenial. **j**, The number of projection targets for every neuron that projects out of V1. **k**, The proportion of cells targeting more than one area, when projection targets that receive projections weaker than the indicated projection strength are ignored. For each neuron, projection strengths are normalized to axon length in the target area receiving the densest innervation. **l**, The fraction of neurons projecting to each of the 18 target areas of V1. **m**, The fraction of neurons innervating a single target area ('dedicated' projection neurons) out of all neurons that innervate that area.

neurons revealed a high degree of diversity in the projections regarding the number and identity of target areas (Fig. 1g and Extended Data Figs 3, 4), whereas this diversity was not clear in the bulk projection data^{20,21} (Fig. 1g, top left).

Almost all layer-2/3 cells projected out of V1 (97%, $n = 30/31$; Fig. 1h) to one or more of 18 target areas in the telencephalon (Fig. 1i), typically innervating nearby cortical areas but occasionally also projecting

to the anterior cingulate cortex, striatum (Extended Data Fig. 1) and amygdala. To mitigate errors arising both from technical noise in atlas registration and from subject-to-subject variability in the boundaries between brain areas, we excluded low-confidence buffer zones of 100 μm around the area boundaries from analysis, and included only those areas that received over 1 mm of axonal input from an individual cell as targets (see Methods). Eighty-five per cent of all projection

patterns appeared only once, highlighting the diversity of long-range projections.

The majority of reconstructed layer-2/3 projection neurons sent axon collaterals to more than one target area (77%, $n = 23/30$), with some targeting up to seven areas (Fig. 1j). Although individual neurons innervated different target areas with different axonal densities, and thus might influence the computations in one area more than another, we found that a large fraction of ‘broadcasting’ cells innervated more than one target with comparable strengths (Fig. 1k). Moreover, the total length of axons scaled with the number of target areas (average length per brain area = 4.6 ± 2.2 mm), such that the innervation density per target was, on average, similar irrespective of how many targets an axon innervated (Extended Data Fig. 6a, b). The innervation in higher visual areas was most dense in layers 2/3 and 5, consistent with recent reports^{26,27}, often recapitulating the pattern of lateral axonal projections of layer-2/3 cells within V1 (Extended Data Fig. 6c–h).

Posterior, postrhinal (POR), lateromedial (LM) and PM visual areas were the most common targets of V1 neurons (Fig. 1l). Even when the analysis was restricted to neurons that projected to at least one of the six nearby cortical visual areas (laterointermediate (LI), LM, AL, PM, anteromedial (AM) or rostrolateral (RL)), we found that half of these neurons projected to two or more of these areas (Extended Data Fig. 7a–e). The fraction of input provided by dedicated projection neurons to any area comprised no more than 25% of the total input (Fig. 1m), and most target areas received no dedicated input. These conclusions were robust to changes in the size of the border exclusion zone between neighbouring areas and the minimum projection strength in the target area (Extended Data Fig. 7f–h). Similar to projections from V1, all seven reconstructed neurons, which had cell bodies that resided in nearby higher visual areas, also projected to more than one target area (Extended Data Fig. 5). Our results thus show that most layer-2/3 neurons distribute information to multiple areas, rather than project to single targets.

Interestingly, the location of the cell body within V1 was predictive of projection target for some recipient areas (Extended Data Fig. 8). Given the retinotopic organization of V1, this suggests that visual information from different parts of visual field may be preferentially distributed to specific target areas, which is consistent with recent findings²⁸.

High-throughput MAPseq tracing

We next investigated whether broadcasting cells choose their cortical target areas independently, or whether they target specific subsets of areas. Although the targeting of different combinations of areas distinguishes individual V1 projection neurons (Fig. 1), their classification into putative sub-types requires a demonstration of higher-order projection structure within the population. We define the higher-order structure in terms of the connection patterns predicted by the per-neuron (first-order) probability of projecting to each target. For example, if the probability of any given neuron projecting to area A is 0.5 and the probability of projecting to area B is also 0.5 then we would expect $P(A \cap B) = P(A) \times P(B) = 0.25$ of all neurons to project to both A and B if the decision to target these areas is independent. Significant deviations from this expectation would indicate the organization of the projections into non-random projection motifs. Investigating the higher-order structure requires large datasets, because, if a sample size of n neurons is required to estimate the first-order probabilities, then a sample size of n^2 is needed to estimate pairwise probabilities with comparable accuracy. Although single-neuron reconstruction provides very high spatial resolution, the tracing of axons remains highly labour intensive despite increases in throughput for data acquisition^{17,29}.

We therefore used a higher-throughput strategy, MAPseq³⁰, to obtain the required number of single-neuron projections for higher-order statistical analysis. In a MAPseq experiment, hundreds or thousands of neurons are labelled uniquely with random RNA sequences (barcodes) by a single injection of a library of barcoded Sindbis viruses (Supplementary Note 2). The barcodes are expressed and then actively

transported into the axonal processes of each labelled neuron, where they can be analysed by high-throughput barcode sequencing after dissection of potential target areas. The abundance of each barcode sequence in each area serves as a measure of the projection strength of the corresponding barcode-labelled neuron. MAPseq simultaneously maps the projections of all labelled neurons of dissected target areas, and therefore enables in-depth analysis of projections to a smaller set of targets.

We used MAPseq to map the projection patterns of 553 neurons from V1 to six higher visual areas—LI, LM, AL, AM, PM and RL—that can be identified reliably by intrinsic signal imaging *in vivo* and dissected *ex vivo* for barcode sequencing (Fig. 2a, b, Extended Data Fig. 9 and Methods). To prevent the virus from spreading from V1 to adjacent areas, we made small focal injections of the MAPseq virus to yield 100–200 traced cells per mouse. Consistent with the analysis of fluorescence-based single-neuron reconstructions restricted to the six higher visual areas (Fig. 2c, left), almost half (44%) of all MAPseq neurons projected to more than one area (Fig. 2c, right). Furthermore, the projection patterns obtained by fluorescence-based tracing were statistically indistinguishable from those obtained by MAPseq (using a bootstrap procedure; see Supplementary Note 3), whereas randomly generated neurons with projection strengths sampled from a uniform distribution were markedly different (Fig. 2d). Therefore, the findings from the MAPseq dataset were consistent with those from the single-neuron tracing dataset.

We first catalogued the diversity of single-neuron projection patterns from V1 to six higher visual areas by unsupervised clustering of the MAPseq dataset (*k*-means clustering with a cosine distance metric). These projection data were best described by eight clusters (Fig. 2e, Extended Data Fig. 10), of which all but one contained cells targeting more than one area. The most common combination of broadcasting neurons involved areas LM and PM, consistent with the fact that a large fraction of neurons targeted these areas and the suggestion of LM²² and PM as integrative hubs of V1 signals, similar to the secondary visual cortex in the monkey (Fig. 2f).

To investigate whether non-random projection motifs existed in the MAPseq dataset, we measured the likelihood of specific bi-, tri- or quadrifurcations and compared them to their expected probabilities (assuming independence between each projection type; Fig. 3a, b). This analysis identified six projection motifs that were significantly over- or underrepresented after a correction for multiple comparisons (Bonferroni adjustment; Fig. 3b, c). Together, these six projection motifs represented 73% of all broadcasting cells that were identified by MAPseq. Therefore the majority of V1 cells projecting to multiple target areas do so in a non-random manner, suggesting that broadcasting motifs reflect several sub-classes of projection neurons for divergent information transfer from V1 to higher visual areas.

The most underrepresented broadcasting motif was the bifurcation between PM and AL (Fig. 3d). These two areas exhibit distinct visual response properties^{13,14} and receive functionally specialized input from V1⁷, consistent with the idea of exclusive projections from V1 into these areas. Moreover, the underrepresented population of neurons, which project to both PM and AL, was further split into two groups according to projection strength; one population primarily innervates PM and another primarily innervates AL (Fig. 3d). A second underrepresented motif is the bifurcation between PM and LM (Fig. 3e). However, in contrast to the PM–AL bifurcation, the detected PM–LM-projection neurons do not clearly separate into two classes. Our findings therefore provide an anatomical substrate for the previously reported functional dichotomy of AL and PM areas, and suggest that a few ‘dedicated’ output channels can co-exist with the prevalence of broadcasting cells that co-innervate multiple targets.

In addition to the two underrepresented projection motifs, we identified four overrepresented motifs, that is, combinations of target areas that receive more shared input from individual V1 neurons than expected from first-order projection statistics (Fig. 3f–h). Cells

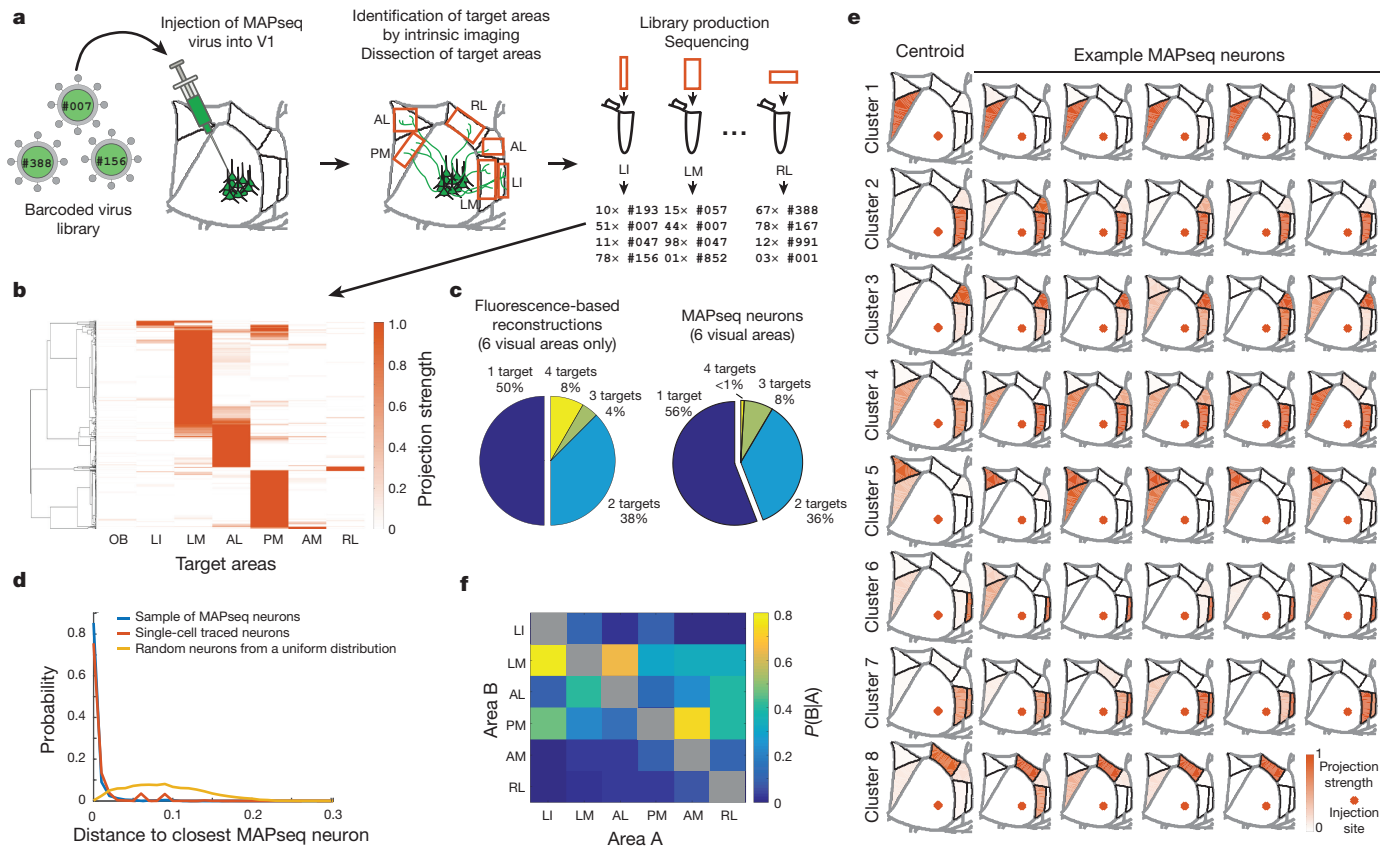


Figure 2 | MAPseq projection mapping reveals a diversity of projection motifs. **a**, Overview of the MAPseq procedure. Six target areas were chosen for analysis: LI, LM, AL, PM, AM and RL. **b**, Projection strength in the six target areas, as well as the olfactory bulb (OB) as a negative control, of 553 neurons mapped using MAPseq. Projection strengths per neuron are defined as the number of barcode copies per area, normalized to the efficiency of sequencing library generation and to the neuron's maximum projection strength ($n = 4$ mice). **c**, Number of projection targets of V1 neurons when considering the six target areas only, based on the fluorescence-based axonal reconstructions (left) or the MAPseq data (right). **d**, Distribution of cosine distances obtained by a bootstrapping procedure (1,000 repeats) between MAPseq neurons (blue), fluorescence-based single-neuron reconstructions and MAPseq neurons (orange), or random neurons (with projection strengths sampled from a uniform

distribution) and MAPseq neurons (yellow). The distance distributions obtained from MAPseq neurons and fluorescence-based single-neuron reconstructions are statistically indistinguishable (Kolmogorov–Smirnov one-sided two-sample test; $P = 0.94$; $\alpha = 0.05$), whereas the distributions obtained from both MAPseq neurons or fluorescence-based reconstructed neurons are statistically different from the distribution obtained using random neurons (Kolmogorov–Smirnov two-sample test; $P < 10^{-3}$; $\alpha = 0.05$). **e**, Centroids and example cells for eight clusters obtained by *k*-means clustering of all MAPseq cells using a cosine distance metric. Target areas are coloured to indicate the projection strength of the plotted neuron. Projection strengths are normalized as in **b**. **f**, The probability of projecting to one area (area A) given that the same neuron is projecting to another area (area B) based on the MAPseq dataset.

that innervated both PM and AM were significantly more abundant than expected by chance (Fig. 3f). Resolving the projection strengths within this motif revealed two subpopulations of neurons, one that innervates PM more than AM, the other innervates both areas with similar strength. Moreover, neurons bifurcating to LM and AL were also highly overrepresented (Fig. 3g) and comprised the most abundant class of broadcasting cells (Fig. 3b). The most significantly over-represented trifurcation motif was the projection to PM, LM and LI, comprising a relatively homogenous population that projects to LM and PM with similar strengths while projecting slightly less strongly to LI (Fig. 3h). Finally, we discovered that trifurcation between PM, AM and RL was overrepresented, but it appeared only rarely in our dataset (Fig. 3b). These motifs did not arise from false negatives (undetected connections) or false positives (Supplementary Note 4 and Extended Data Fig. 2f).

These projectional data have implications for the categorization of higher visual areas into putative streams of visual processing in mouse neocortex. Areas AL and PM on the one hand, and LM and LI on the other, have been suggested to belong to dorsal and ventral processing streams in the mouse visual system, respectively^{31–33}. Given that these areas receive a high degree of shared input (for example, LM–PM

bifurcation, which was still abundant even though it was underrepresented; AL–LM bifurcation; PM–LM–LI trifurcation), such a distinction is unlikely to originate as a result of segregated V1 input into these areas.

Discussion

In summary, our results reveal some of the principles by which single neurons in one cortical area distribute information to downstream target areas. Almost all layer-2/3 pyramidal cells projected outside of V1, indicating that V1 neurons concurrently participate in local and distal computations. We found that the single-neuron projections outside V1 were highly diverse, innervating up to seven targets, predominantly in specific, non-random combinations (Extended Data Fig. 10g, f). These results suggest a functional specialization of subpopulations of projection cells beyond ‘one neuron–one target area’ mapping.

The fraction of neurons in V1 that broadcast information to multiple targets is considerably greater than has previously been indicated using retrograde tracing methods^{2,5,16}. This difference is unlikely to be caused by differences in the sensitivity with which these approaches detect the projection patterns of individual cells. Instead, anterograde tracing maps projections to many or all targets simultaneously, whereas

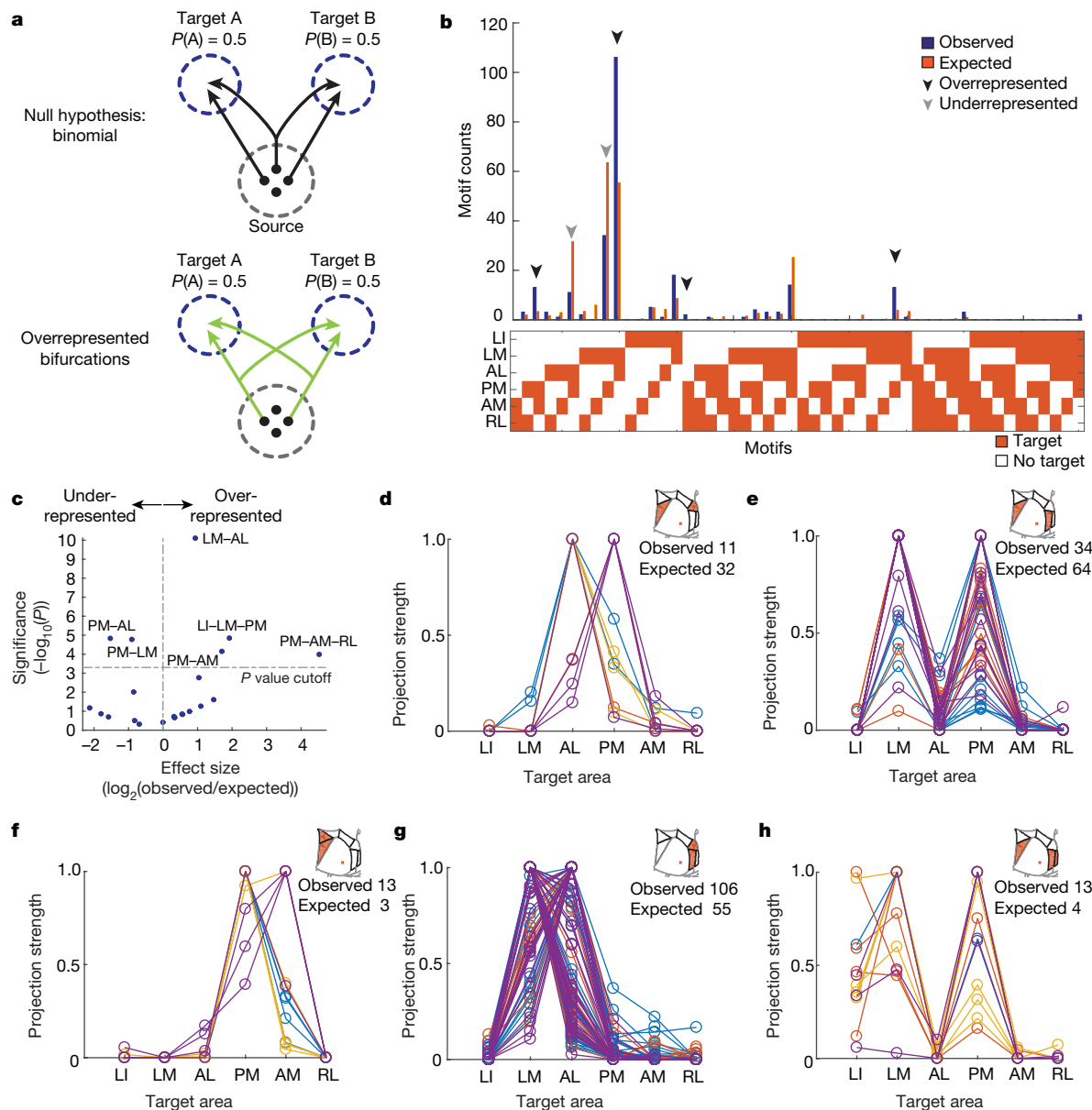


Figure 3 | Over- and underrepresented projection motifs of neurons in V1. **a**, The null hypothesis of independent projections to two target areas (top) and an example deviation (overrepresented bifurcation) from the null hypothesis (bottom). **b**, The observed and expected abundance of all possible bi-, tri- and quadrifurcation motifs in the MAPseq dataset. Significantly over- or underrepresented motifs, based on a binomial test with Bonferroni correction (see Methods), are indicated by black and grey arrowheads, respectively. $n = 553$ neurons from four mice. **c**, Statistical

significance of over- and underrepresented broadcasting motifs and associated effect sizes, based on a binomial test with Bonferroni correction (see Methods). $n = 553$ neurons from four mice. **d–h**, The projection strengths of the individual neurons (one per line) giving rise to the six underrepresented (**d**, **e**) and overrepresented (**f–h**) projection motifs. For each neuron, the projection strength in each target area is normalized to the neuron's maximum projection strength. Lines of the same colour represent neurons mapped in the same brain ($n = 4$ mice).

retrograde tracing typically analyses only two or three potential target sites at a time. Because the fraction of neurons projecting to any pair of targets selected for retrograde tracing is relatively low (typically $< 10\%$), most neurons will not be doubly labelled in any given experiment; only by sampling many potential targets in a single experiment can the true prevalence of broadcasting be uncovered. Indeed, if we simulate double retrograde tracing based on our MAPseq results, the fractions of bifurcating neurons are comparable to those observed when using retrograde methods in primates^{2,5,16,18} (Supplementary Table 1).

We speculate that dedicated projection neurons—which comprise the minority of neurons in V1—convey specialized visual information that is tailored to their target area, as has previously been suggested^{6–11}. Indeed, the most underrepresented projection motif from V1, the PM–AL bifurcation, innervates two target areas with distinct preferences

for visual features^{13,14}. By contrast, we suggest that the majority of cells encode information that is shared and in a form that is suitable for generating visual representations or multimodal associations across subsets of areas. Indeed, those target areas that are preferentially co-innervated by broadcasting neurons appear to have more similar visual response properties^{13,14}. Broadcasting cells may also coordinate activity among the subset of areas that they co-innervate, thus providing a signal that links different processing streams. The divergent nature of signal transmission from a primary sensory cortex to its targets may therefore help to constrain models of hierarchical sensory processing. The existence of distinct projection motifs that either avoid or favour subsets of target areas suggests that sub-types of intracortical projection neurons exist and raises the question of how these specific, long-range connectivity patterns are established during development.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 June 2017; accepted 31 January 2018.

Published online 28 March 2018.

- Harris, K. D. & Shepherd, G. M. G. The neocortical circuit: themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
- Nakamura, H., Gattass, R., Desimone, R. & Ungerleider, L. G. The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J. Neurosci.* **13**, 3681–3691 (1993).
- Segraves, M. A. & Innocenti, G. M. Comparison of the distributions of ipsilaterally and contralaterally projecting corticocortical neurons in cat visual cortex using two fluorescent tracers. *J. Neurosci.* **5**, 2107–2118 (1985).
- Rockland, K. S. Collateral branching of long-distance cortical projections in monkey. *J. Comp. Neurol.* **521**, 4112–4123 (2013).
- Sincich, L. C. & Horton, J. C. Independent projection streams from macaque striate cortex to the second visual area and middle temporal area. *J. Neurosci.* **23**, 5684–5692 (2003).
- Yamashita, T. *et al.* Membrane potential dynamics of neocortical projection neurons driving target-specific signals. *Neuron* **80**, 1477–1490 (2013).
- Glickfeld, L. L., Andermann, M. L., Bonin, V. & Reid, R. C. Cortico-cortical projections in mouse visual cortex are functionally target specific. *Nat. Neurosci.* **16**, 219–226 (2013).
- Sato, T. R. & Svoboda, K. The functional properties of barrel cortex neurons projecting to the primary motor cortex. *J. Neurosci.* **30**, 4256–4260 (2010).
- Chen, J. L., Carta, S., Soldado-Magraner, J., Schneider, B. L. & Helmchen, F. Behaviour-dependent recruitment of long-range projection neurons in somatosensory cortex. *Nature* **499**, 336–340 (2013).
- Yamashita, T. & Petersen, C. C. H. Target-specific membrane potential dynamics of neocortical projection neurons during goal-directed behavior. *eLife* **5**, e15798 (2016).
- Movshon, J. A. & Newsome, W. T. Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *J. Neurosci.* **16**, 7733–7741 (1996).
- Nassi, J. J. & Callaway, E. M. Parallel processing strategies of the primate visual system. *Nat. Rev. Neurosci.* **10**, 360–372 (2009).
- Andermann, M. L., Kerlin, A. M., Roumis, D. K., Glickfeld, L. L. & Reid, R. C. Functional specialization of mouse higher visual cortical areas. *Neuron* **72**, 1025–1039 (2011).
- Marshall, J. H. H., Garrett, M. E. E., Nauhaus, I. & Callaway, E. M. Functional specialization of seven mouse visual cortical areas. *Neuron* **72**, 1040–1054 (2011).
- Massé, I. O., Régnier, P. & Boire, D. in *Axons and Brain Architecture* (ed. Rockland, K. S.) Ch. 5, 93–116 (Academic, 2016).
- Bullier, J. & Kennedy, H. Axonal bifurcation in the visual system. *Trends Neurosci.* **10**, 205–210 (1987).
- Economo, M. N. *et al.* A platform for brain-wide imaging and reconstruction of individual neurons. *eLife* **5**, e10566 (2016).
- Vogt Weisenhorn, D. M., Illing, R. B. & Spatz, W. B. Morphology and connections of neurons in area 17 projecting to the extrastriate areas MT and 19DM and to the superior colliculus in the monkey *Callithrix jacchus*. *J. Comp. Neurol.* **362**, 233–255 (1995).
- Ding, S.-L., Van Hoesen, G. & Rockland, K. S. Inferior parietal lobule projections to the presubiculum and neighboring ventromedial temporal cortical areas. *J. Comp. Neurol.* **425**, 510–530 (2000).
- Zingg, B. *et al.* Neural networks of the mouse neocortex. *Cell* **156**, 1096–1111 (2014).
- Oh, S. W. *et al.* A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
- Wang, Q. & Burkhalter, A. Area map of mouse visual cortex. *J. Comp. Neurol.* **502**, 339–357 (2007).
- Ragan, T. *et al.* Serial two-photon tomography for automated *ex vivo* mouse brain imaging. *Nat. Methods* **9**, 255–258 (2012).
- Osten, P. & Margrie, T. W. Mapping brain circuitry with a light microscope. *Nat. Methods* **10**, 515–523 (2013).
- Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
- D'Souza, R. D., Meier, A. M., Bista, P., Wang, Q. & Burkhalter, A. Recruitment of inhibition and excitation across mouse visual cortex depends on the hierarchy of interconnecting areas. *eLife* **5**, e19332 (2016).
- Yang, W., Carrasquillo, Y., Hooks, B. M., Nerbonne, J. M. & Burkhalter, A. Distinct balance of excitation and inhibition in an interareal feedforward and feedback circuit of mouse visual cortex. *J. Neurosci.* **33**, 17373–17384 (2013).
- Zhuang, J. *et al.* An extended retinotopic map of mouse cortex. *eLife* **6**, e18372 (2017).
- Gong, H. *et al.* High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nat. Commun.* **7**, 12142 (2016).
- Kebschull, J. M. *et al.* High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron* **91**, 975–987 (2016).
- Wang, Q., Sporns, O. & Burkhalter, A. Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. *J. Neurosci.* **32**, 4386–4399 (2012).
- Smith, I. T., Townsend, L. B., Huh, R., Zhu, H. & Smith, S. L. Stream-dependent development of higher visual cortical areas. *Nat. Neurosci.* **20**, 200–208 (2017).
- Murakami, T., Matsui, T. & Ohki, K. Functional segregation and development of mouse higher visual areas. *J. Neurosci.* **37**, 9424–9437 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Juavinett, L. Huang, S. Hofer and P. Znamenskiy for comments on the manuscript. This study was funded by National Institutes of Health (5R01NS073129 and 5R01DA036913 to A.M.Z.); Brain Research Foundation (BRF-SIA-2014-03 to A.M.Z.); IARPA (MICrONS D16PC0008 to A.M.Z.); Simons Foundation (382793/SIMONS to A.M.Z.); Paul Allen Distinguished Investigator Award (to A.M.Z.); PhD fellowship from the Boehringer Ingelheim Fonds (to J.M.K.); PhD fellowship from the Genentech Foundation (to J.M.K.); National Natural Science Foundation of China (NSFC 31600847 to Y.H.); European Research Council (NeuroVision 616509 to T.D.M.-F.), and Swiss National Science Foundation (SNSF 31003A_169802 to T.D.M.-F.).

Author Contributions Y.H. generated the dataset for fluorescence-based axonal tracing. D.C. and Y.H. traced the cells. R.A.A.C. analysed the serial two-photon imaging data and axonal projection patterns. J.M.K. and F.I. collected the MAPseq dataset. J.M.K. and A.M.Z. performed the analysis of projection patterns. J.M.K., T.D.M.-F. and A.M.Z. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.M.Z. (zador@cshl.edu) or T.D.M.-F. (t.mrsic-flogel@ucl.ac.uk).

Reviewer Information *Nature* thanks M. Helmstaedter, O. Sporns and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Mice. The anatomical single-cell tracing experiments were conducted at The Biozentrum, University of Basel, Switzerland. We obtained licenses and performed all experimental procedures in accordance with Basel Canton animal welfare guidelines using both male and female adult (>8 weeks of age) C57BL/6 mice. Detailed protocols and all software are available at <http://mouse.vision/han2017>.

Fluorescence-based single-neuron tracing. *Two-photon guided single-cell electroporation.* We performed surgery as previously described³⁴. In brief, we anaesthetized mice with a mixture of fentanyl (0.05 mg kg⁻¹), midazolam (5 mg kg⁻¹) and medetomidine (0.5 mg kg⁻¹), and maintained stable anaesthesia by isoflurane (0.5% in O₂). We performed all electroporation experiments on a custom linear scanning two-photon microscope, equipped to image both a green and a red channel and running ScanImage 5.1³⁵. For electroporation, we used a patch pipette (12–16 M Ω) filled with plasmid DNA (pCAG-eGFP (Addgene, 11150) or pAAV-EF1a-eGFP-WPRE (gift from B. Roska; sequence file can be found in the Supplementary Information, 100 ng μ l⁻¹) and AlexaFluor 488 (50 μ M) in intracellular solution, and delivered electroporation pulses (100 Hz, -14 V, 0.5-ms duration for 1 s) with an Axoporation 800A (Molecular Probes) when pushed against a target cell. We verified successful electroporation by dye filling of the cell body, and then sealed the skull with a chronic window using 1.5% agarose in HEPES-buffered artificial cerebrospinal fluid and a cover slip. We confirmed plasmid expression two days after electroporation by visualization of GFP epifluorescence through the chronic imaging window. Three to ten days after electroporation, we transcardially perfused anaesthetized mice with 10 ml 0.9% NaCl followed by 50 ml 4% paraformaldehyde in 0.1 M phosphate buffer (pH 7.4). We removed the brains from the skull and post-fixed them in 4% paraformaldehyde overnight at 4°C. We then stored the fixed brains in PBS at 4°C until imaging with serial-section two-photon tomography.

Serial-section two-photon tomography. We embedded the fixed brains in 5% oxidized agarose (derived from type-I agarose (Sigma-Aldrich)) and covalently cross-linked the brain to the agarose by incubation in an excess of 0.5–1% sodium borohydride (NaBH_4 , Sigma-Aldrich) in 0.05 M sodium borate buffer overnight at 4°C. We then imaged embedded brains using a TissueVision two-photon scanning microscope^{23,36}, which cut physical sections of the entire brain every 50 μm coronally, and acquired optical sections every 10 μm in two channels (green channel: 500–560 nm; red channel: 560–650 nm) using 940-nm excitation laser light (Mai Tai eHP, Spectraphysics). Each imaged section is formed from overlapping 800 \times 800- μm ‘tiles’. We imaged with a resolution of 1 μm in x and y and measured an axial point spread function of $\sim 5 \mu\text{m}$ FWHM (full width at half maximum) using ScanImage 5.1.

Image processing and cell tracing. We stitched raw image tiles using custom MATLAB-based software, StitchIt. StitchIt applies illumination correction based on the average tiles for each channel and optical plane, and subsequently stitches the illumination-corrected tiles from the entire brain. We then navigated through the stitched brain space using MaSIV (<https://github.com/alexanderbrown/masiv>), a MATLAB-based viewer for very large 3D images, and traced axons using a custom, manual neurite-tracing extension for MaSIV. The tracer was not blinded, as no comparison across experimental conditions was performed. No power calculations were performed.

To assign each voxel of the imaged brains to a brain area, we segmented each brain using areas defined by the Allen Reference Atlas (Common Coordinate Framework v.3, © 2015 Allen Institute for Brain Science, Allen Brain Atlas API, available from <http://brain-map.org/api/index.html>), after smoothing with a single pass of a Gaussian kernel with an s.d. of 0.5 using the Nifty ‘seg-maths’ tool as described previously³⁷. In brief, we downsampled one imaging channel to a voxel size of 25 μm and converted it to MHD format using StitchIt. We then registered the volume to the average template brain of the Allen Reference Atlas using Elastix³⁸ by applying rigid affine transformation followed by non-rigid deformation with parameters as previously described^{39,40}. We examined registration quality using a custom Python/PyQt5 application, Lasagna, which overlays the Allen template brain and the registered sample brain and is extendable to allow the overlay of traced cells, or the overlay of area borders from the Allen Reference Atlas onto a downsampled brain. To transform the traced cells into Allen Reference Atlas space (sample to the Allen Reference Atlas), we calculated the inverse transform to the one calculated by Elastix (Allen Reference Atlas to sample) and applied this to the traced points.

Analysis of traced neurons. To prevent potential incomplete filling of neurons from biasing the results of our analyses, we excluded cells with non-arborizing primary branches in the ipsilateral hemisphere from the analysis. Out of a total of 71 traced cells, we excluded 28 cells that exhibited abrupt, non-callosal terminations, as well as 5 cells that were back-labelled from the striatum, thus restricting our analysis to ipsilateral projection patterns of 31 cells in V1 and 7 in other higher visual areas.

Moreover, axonal branches terminating contralaterally or after entering the corpus callosum were considered as callosal terminations and were included in the analysis (see also ref. 6). We calculated the first-order projection statistics only using the cells registered in the Allen Reference Atlas that satisfied these criteria. To reduce any artefacts associated with registration in the Allen Reference Atlas or individual brain variability in boundaries between brain areas, we excluded any axon within 50 μm from any brain area boundary from the analysis. We then calculated the projection strength of each neuron to each area as the total length of axon of that neuron in an area. To determine the number of projection targets for every cell, we used a minimum projection strength of 1-mm axon length per target area.

MAPseq. MAPseq sample processing. To define the V1 injection site and target higher visual areas (LI, LM, AL, PM, AM and RL), we used optical imaging of intrinsic signals as previously described^{13,41}. In brief, we first implanted a customized head plate and then thinned the skull to increase its transparency. After 2–3 days of recovery, we sedated the mice (chlorprothixene, 0.7 mg kg⁻¹) and lightly anaesthetized them with isoflurane (0.5–1.5% in O₂), delivered via a nose cone. We illuminated the visual cortex with 700-nm light that was split from an LED source into two light guides, performing imaging with a tandem lens macroscope focused 250–500 μm below the cortical surface and a bandpass filter centred at 700 nm with 10-nm bandwidth (67905; Edmund Optics). We acquired images at 6.25 Hz with a 12-bit CCD camera (1300QF; VDS Vosskühler), frame grabber (PCI-1422; National Instruments) and custom software written in LabVIEW (National Instruments). We visually stimulated the contralateral eye of mice with a monitor placed at a distance of 21 cm and presented 25–35° patches of 100% contrast square wave gratings with a temporal frequency of 4 Hz and a spatial frequency of 0.02 cycles per degree for 2 s followed by 5 s of grey screen (mean luminance of 46 candela per m²). To establish a coarse retinotopy of the targeted area, we alternated the position of the patches: we used two different elevations (approximately 0 and 20°) and two different azimuths (approximately 60 and 90°); at each position, we acquired at least 17 trials. We obtained intrinsic signal maps by averaging the responses during the stimulation time using ImageJ (National Institute of Mental Health, NIH) and mapping the location of the estimated spots of activation onto a previously acquired blood vessel picture.

We then pressure-injected (Picospritzer III, Parker) 100 nl of 1×10^{10} genome copies ml^{-1} barcoded MAPseq Sindbis virus³⁰ with a diversity of $>8 \times 10^6$ different barcode sequences unilaterally at a depth of 100–200 μm from the brain surface into V1 of four 8–10-week-old C57BL/6 female mice. In addition, we labelled the six higher visual areas by placing a DiI-coated micropipette into retinotopically matched positions according to intrinsic signal maps. For this, we allowed 2–5 μl of 2.5 mg ml^{-1} DiI (Invitrogen D3911) in ethanol solution to dry on the outside of a pulled micropipette tip until some DiI crystals were visible. Mice were euthanized 44–48 h after injection by decapitation, and their brain immediately extracted and flash-frozen on dry ice.

We cut 180- μm thick coronal sections using a cryostat at -10°C blade and sample holder temperature, and melted each slice onto a clean microscope slide before rapidly freezing it on dry ice again. We then dissected each target area and the injection site using cold scalpels while keeping the brain sections frozen on a metal block cooled to approximately -20°C in a freezing 2.25 M CaCl_2 bath⁴². During dissection, we identified each dissected area using a fluorescent dissection microscope to visualize viral GFP expression and DiI stabs labelling each target area (Extended Data Fig. 7). Throughout the procedure, we took care to avoid sample cross-contamination by never reusing tools or blades applied to different areas and changing gloves between samples. To measure noise introduced by contamination, we collected samples of the olfactory bulb from each brain, which served as a negative control.

We then processed the dissected samples for sequencing mostly as previously described³⁰, but pooling all samples after first-strand cDNA synthesis. In brief, we extracted total RNA from each sample using TRIzol reagent (Thermo Fisher) according to the manufacturer's instructions. We mixed the sample RNA with spike-in RNA (obtained by *in vitro* transcription of a double-stranded ultramer with sequence 5'-GTCATGATCATAATACGACTCACTATAG GGGACGAGCTGTACAAGTAAACGCGTAATGATACGGCGACCACCGAGA TCTACACTCTTTCCCTACACGACGCTCTCCGATCTNNNNNNNNNNNNNN NNNNNNNNNNNNNNATCAGTCATCGGAGCGGCCGCTACCTAATTGCCG CCGTGAGGTACGACACCGCATGCTGTACA-3' (IDT)³⁰) and reverse transcribed the RNA mixture using a gene specific primer 5'-CTTGGCACCCGA GAATTCANNNNNNNNNNNNNNXXXXXXXXXGTACAGCTAGCGGTGGT CG-3', where X₈ is one of >300 true-seq-like sample-specific identifiers and N₁₂ is the unique molecular identifier, and SuperscriptIV Reverse Transcriptase (Thermo Fisher) according to the manufacturer's instructions. We then pooled all first-strand cDNAs, purified them using SPRI beads (Beckman

Coulter) and produced double-stranded cDNA as previously described⁴³. We then treated the samples using Exonuclease I (NEB) and performed two rounds of nested PCR using primers 5'-CTCGGCATGGACGAGCTGTA-3' and 5'-CAAGCAGAAGACGGCATACGAGATCGTGTGACTGGAGTTCCTTGGCACCAGGAATCCA-3' for the first PCR and primers 5'-AATGATACGGCGACCACCGA-3' and 5'-CAAGCAGAAGACGGCATACGA-3' for the second PCR using Accuprime Pfx polymerase (Thermo Fisher). Finally, we gel-extracted the resulting PCR amplicons using Qiagen MinElute Gel extraction kit according to the manufacturer's instructions and sequenced the library on a Illumina NextSeq500 high-output run at paired-end 36 using the SBS3T sequencing primer for paired-end 1 and the Illumina small RNA sequencing primer 2 for paired-end 2.

MAPseq data analysis. On the basis of the sequencing results, we constructed a barcode matrix M of size of (number of barcodes) \times (number of dissected areas) with entry $M_{i,j}$ representing the absolute counts of barcode i in area j as previously described³⁰. We de-multiplexed the sequencing results, extracted the absolute counts of each barcode in each sample based on the UMI sequence and error-corrected the barcode sequences, before matching barcode sequences to the virus library and constructing matrix M by matching barcode sequences across areas. We then filtered the barcode matrix for 'high-confidence' cell bodies inside the dissected area of V1 by requiring a minimum of 10 counts in at least one target area, an at least tenfold difference between the cell body location in V1 and the most abundant target area in data normalized to the efficiency of library production as measured by the amount of recovered spike-in RNA counts, and an absolute minimum barcode count of 300 in V1. We then normalized the raw barcode counts in each area to the relative spike-in RNA recovery to the olfactory bulb sample, merged the results from all four processed brains into a single barcode matrix and used this matrix for all further analysis.

To determine whether a particular neuron projected to any given target area, we chose a conservative threshold of at least 5 barcode counts, based on the highest level of barcode expression in the olfactory bulb negative control sample.

Calculation of statistical significance of projection motifs. To calculate the statistical significance of broadcasting projection motifs, we compared against the simplest model in which we assumed that each neuron projected to each area independently. To generate predictions of this model, we first estimated the probability of projecting to each area, assuming independent projections. We define the probability $P(A_i)$ that a given neuron projects to the i th area A_i as

$$P(A_i) = \frac{N_{A_i}}{N_{\text{total}}}$$

in which N_{A_i} is the number of neurons in the sample that project to area A_i , $i = 1 \dots k$ for k analysed target areas, and N_{total} is the total number of neurons in the sample.

In our MAPseq experiments, we do not have direct access to N_{total} , since for technical reasons we only include neurons that have at least one projection among the dissected targets. Because, in principle, some neurons might project to none of the areas dissected (see Fig. 3a), failure to include these would lead to an underestimation of N_{total} . However, assuming independence of projections, we can infer N_{total} from the available measurements.

To estimate N_{total} , we first observe that

$$P(\text{project to at least one area}) + P(\text{project to no area}) = 1$$

$$\Leftrightarrow \frac{N_{\text{obs}}}{N_{\text{total}}} + \prod_{j=1}^k \left(1 - \frac{N_{A_j}}{N_{\text{total}}}\right) = 1$$

where N_{obs} is the total number of neurons observed to project to at least one area. For $k=6$ areas, we can expand this expression to

$$\begin{aligned} & \left(N_{\text{obs}} - \sum_{j=1}^6 N_{A_j} \right) N_{\text{total}}^5 + \sum_{i=1}^6 \sum_{j=1}^6 N_{A_i} N_{A_j} N_{\text{total}}^4 \\ & - \sum_{i=1}^6 \sum_{j=1}^6 \sum_{k=1}^6 N_{A_i} N_{A_j} N_{A_k} N_{\text{total}}^3 \\ & + \sum_{i=1}^6 \sum_{j=1}^6 \sum_{k=1}^6 \sum_{l=1}^6 N_{A_i} N_{A_j} N_{A_k} N_{A_l} N_{\text{total}}^2 \\ & - \sum_{i=1}^6 \sum_{j=1}^6 \sum_{k=1}^6 \sum_{l=1}^6 \sum_{m=1}^6 N_{A_i} N_{A_j} N_{A_k} N_{A_l} N_{A_m} N_{\text{total}} + \prod_{i=1}^6 N_{A_i} = 0 \end{aligned}$$

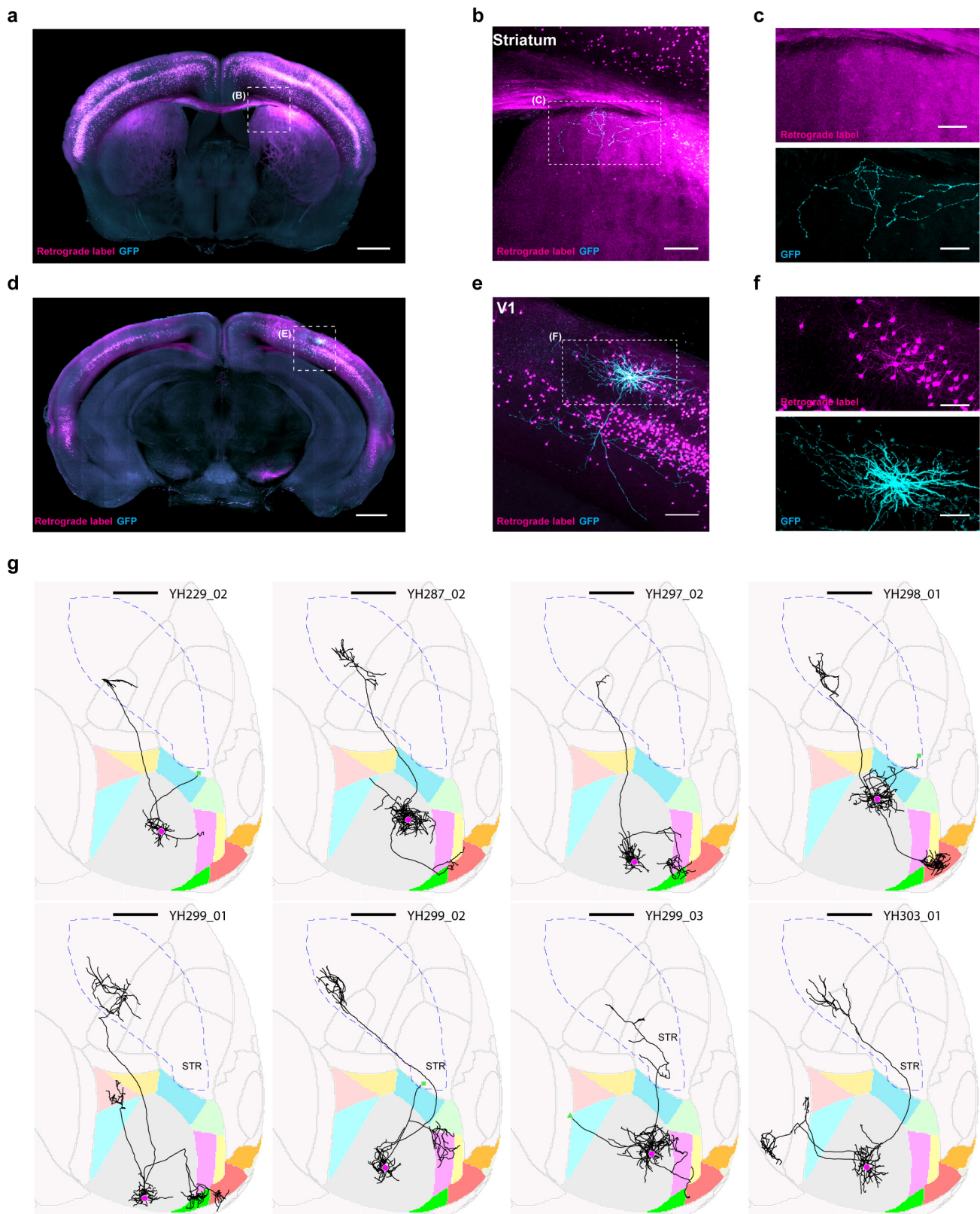
Noting that this is a quintic equation in N_{total} , we can use a root finder to solve for N_{total} numerically, and use the result to calculate $P(A_i)$.

Using the derived N_{total} and $P(A_i)$, we can calculate the P value for every possible broadcasting motif by calculating the value of the binomial cumulative distribution function, for a total of N_{total} tries, the empirical number of observed counts (successes), and $P(\text{motif})$ assuming independent projections. We calculated the P value of all possible bi-, tri- and quadrifurcations, and determined significantly over- or underrepresented broadcasting motifs at a significance threshold of $\alpha = 0.05$ after Bonferroni correction.

Code availability. All software is available at <http://mouse.vision/han2017>.

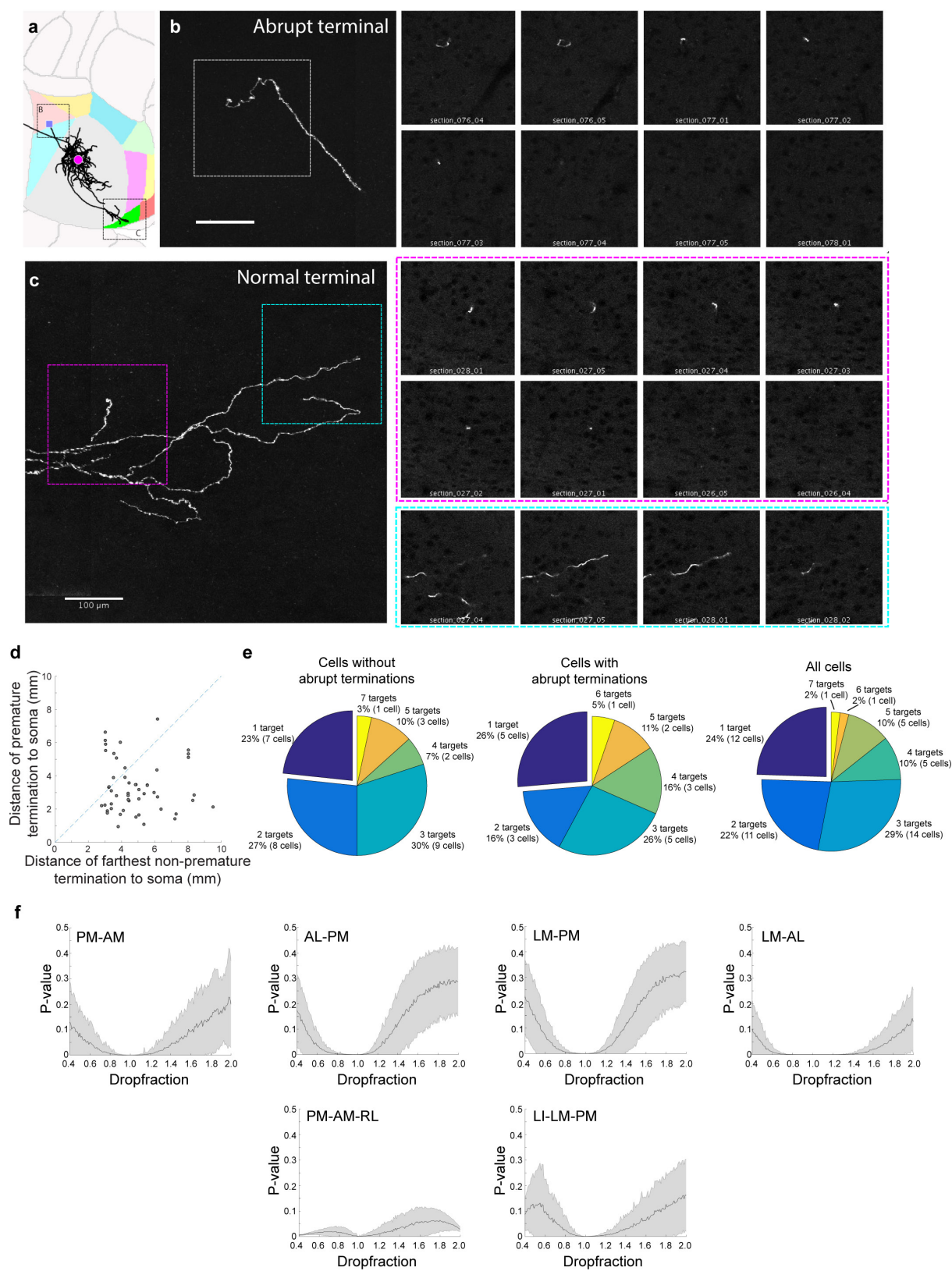
Data availability. All sequencing data are publicly available in the Sequence Read Archive under accession numbers SRR5274845 (ZL097 for mouse 4 and mouse 5) and SRR5274844 (ZL102 for mouse 6 and mouse 7). All single-cell tracing results are accessible at <http://mouse.vision/han2017> and on neuromorpho at http://neuromorpho.org/NeuroMorpho_ArchiveLinkout.jsp?ARCHIVE=Han_et_al.

34. Pecka, M., Han, Y., Sader, E. & Mrsic-Flogel, T. D. Experience-dependent specialization of receptive field surround for selective coding of natural scenes. *Neuron* **84**, 457–469 (2014).
35. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
36. Mayerich, D., Abbott, L. & McCormick, B. Knife-edge scanning microscopy for imaging and reconstruction of three-dimensional anatomical structures of the mouse brain. *J. Microsc.* **231**, 134–143 (2008).
37. Niedworok, C. J. *et al.* aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. *Nat. Commun.* **7**, 11879 (2016).
38. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29**, 196–205 (2010).
39. Kim, Y. *et al.* Whole-brain mapping of neuronal activity in the learned helplessness model of depression. *Front. Neural Circuits* **10**, 3 (2016).
40. Renier, N. *et al.* Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* **165**, 1789–1802 (2016).
41. Roth, M. M. *et al.* Thalamic nuclei convey diverse contextual information to layer 1 of visual cortex. *Nat. Neurosci.* **19**, 299–307 (2016).
42. Bryan, W. P. & Byrne, R. H. A calcium chloride solution, dry-ice, low temperature bath. *J. Chem. Educ.* **47**, 361 (1970).
43. Morris, J., Singh, J. M. & Eberwine, J. H. Transcriptome analysis of single cells. *J. Vis. Exp.* **50**, e2634 (2011).



Extended Data Figure 1 | Single-neuron tracing protocol efficiently fills axons projecting to the ipsilateral striatum. We retrogradely labelled striatum-projecting cells by stereotactically injecting cholera toxin subunit B conjugated to AlexaFluor 594 or PRV-Cre into the visual striatum of wild-type mice or tdTomato-reporter mice (Ai14, Jax), respectively (magenta). With visual guidance of two-photon microscopy, we electroporated single retrogradely labelled cells in V1 with a GFP-expressing plasmid (cyan). **a**, Coronal, maximum intensity projections of visual striatum. Scale bar, 1 mm. **b**, Higher magnification image of the visual striatum. Scale bar, 0.2 mm. **c**, Single-channel images of the same

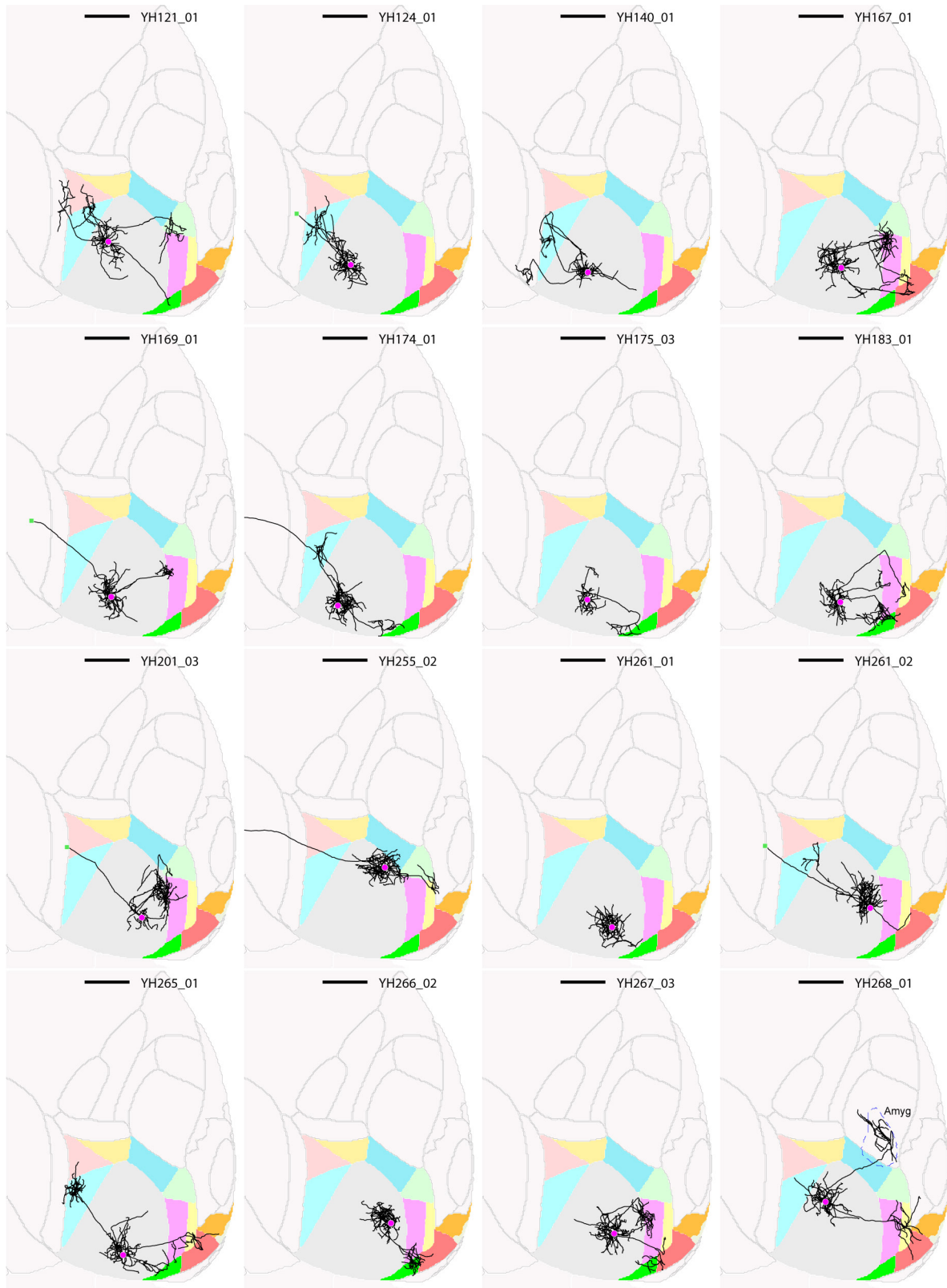
axonal arborization as in **b**, **d**. Coronal maximum intensity projection containing V1. Scale bar, 1 mm. **e**, Higher magnification image of V1. Scale bar, 0.2 mm. **f**, Single-channel images of V1. Scale bars, 0.2 mm. **g**, Horizontal projections in the Allen Reference Atlas space of eight retrogradely labelled and electroporated cells. Cell ID numbers are indicated at the top right of each image. Scale bars, 1 mm. Note that one additional cell was retrogradely labelled and electroporated, which revealed its axonal projection to the striatum, but it is not shown because the brain was too distorted to allow accurate registration to the Allen Reference Atlas.



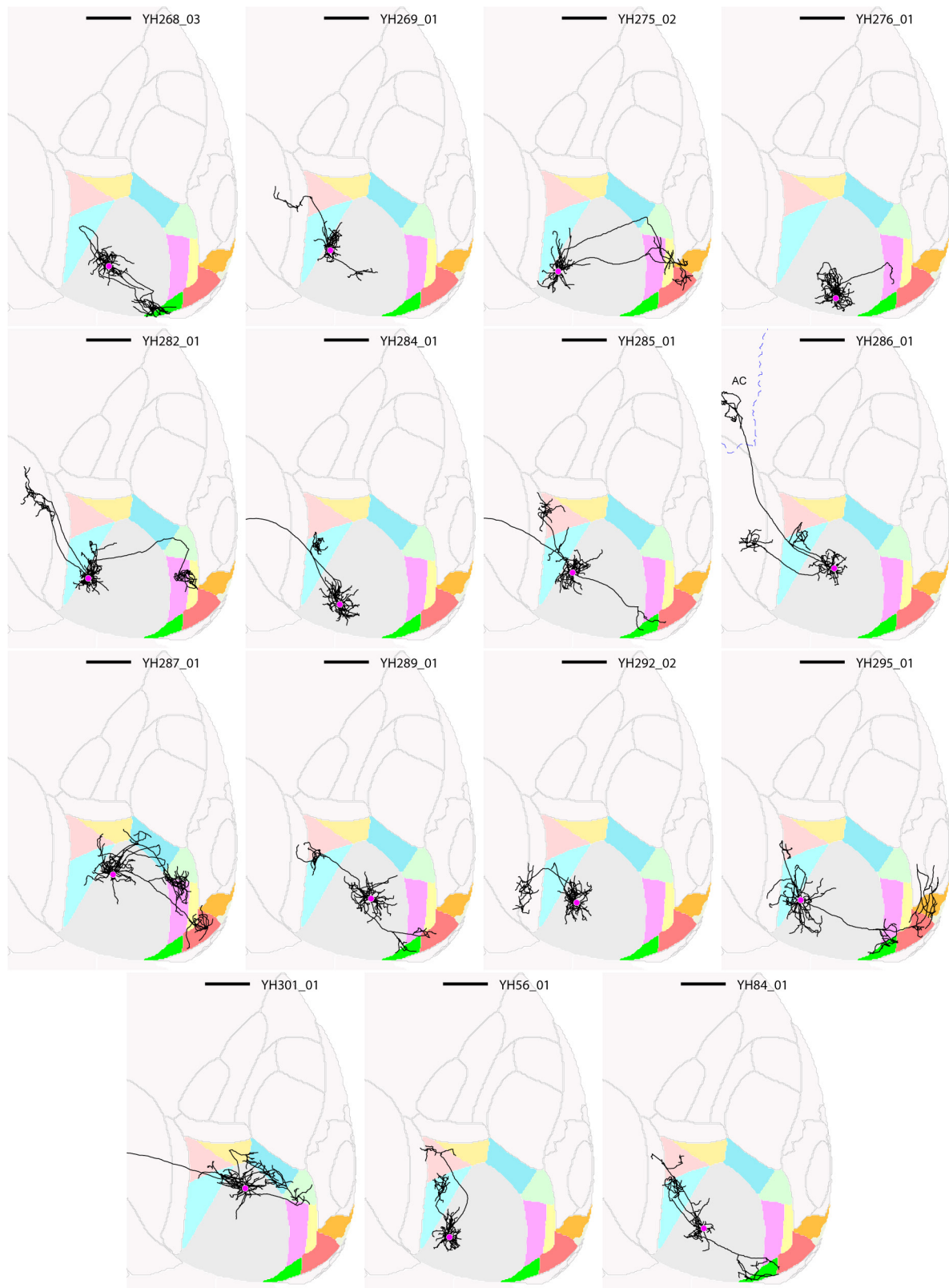
Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Some axonal branches terminate abruptly without arborizing, whereas other branches of the same neuron arborize extensively within different target areas and appear to be completely filled. **a**, Horizontal view of a representative cell in the Allen Reference Atlas space. The abrupt termination is labelled with a purple square. $n = 28$ abruptly terminating cells. **b**, The abrupt termination of the example cell shown as a maximum z projection (left) and in the individual z sections (right). **c**, Two normal terminations of the same cell, shown as a maximum z projection (left) and in two colour-coded series of z sections (right). **d**, Distance of abrupt termination from cell body versus distance of furthest regular termination of the same cell. Dashed line indicates the

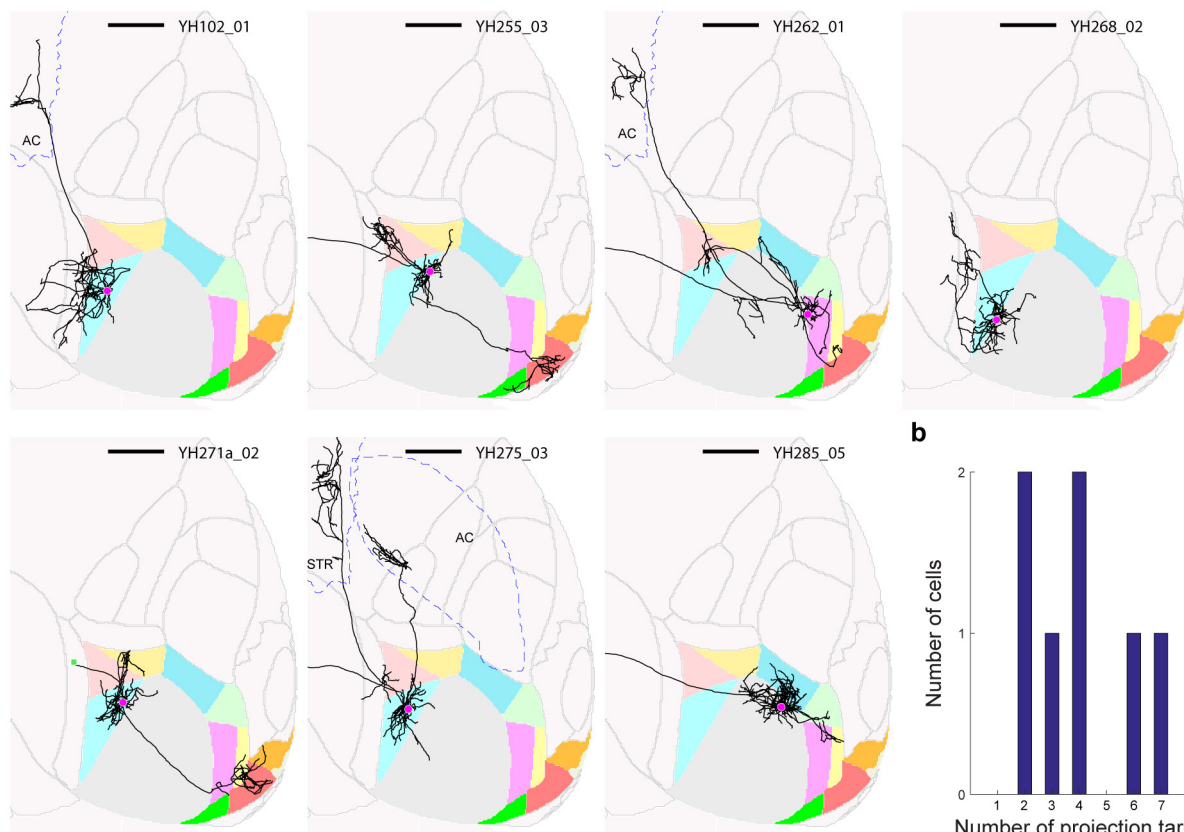
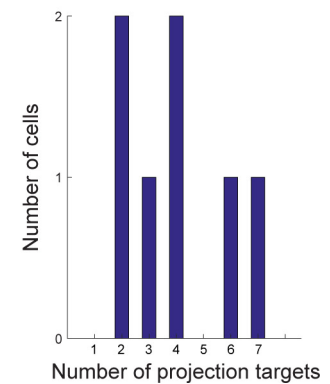
unity line. **e**, The distribution of target numbers of all projection neurons without abrupt terminations (as shown in the main figures; left), of projection cells with abrupt terminations (middle) and of all projection neurons (no abrupt terminations and abrupt terminations; right). **f**, To test the effect of false negatives on our analyses, we simulated the random loss or gain of projections from the MAPseq dataset, while maintaining overall area projection probabilities. $n = 553$ neurons; 400 repeats. P values based on a binomial test for all six projection motifs determined as significantly over- or underrepresented in our dataset are plotted after removing (dropfraction < 1) or adding (dropfraction > 1) connections. Mean (black line) and s.d. (shaded area) are indicated.



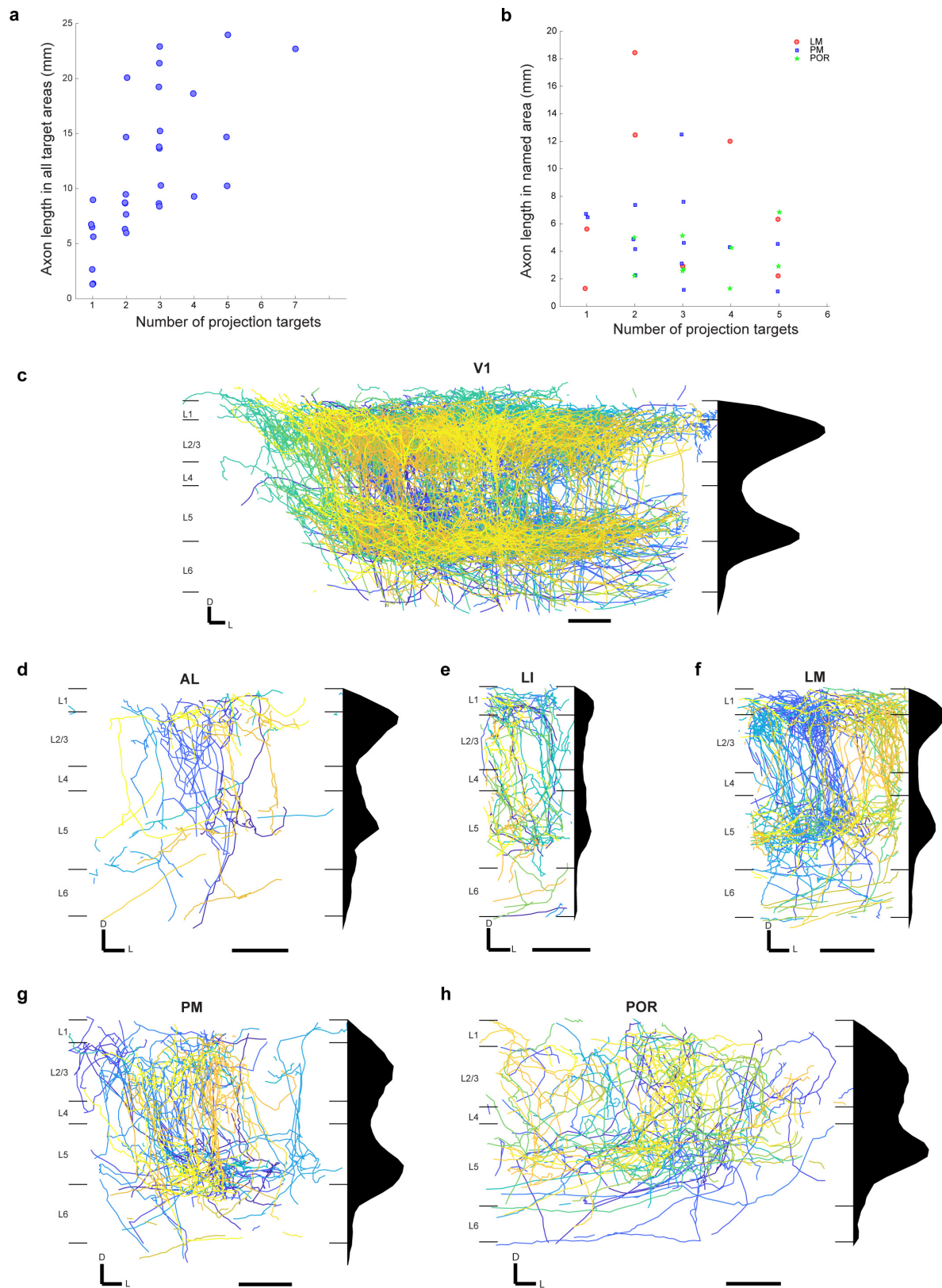
Extended Data Figure 3 | Images of traced layer-2/3 V1 neurons. Horizontal views of the Allen Reference Atlas space are shown, and cell ID numbers are indicated at the top right of each image. Scale bars, 1 mm.



Extended Data Figure 4 | Images of traced layer-2/3 V1 neurons, continued. Horizontal views of the Allen Reference Atlas space are shown, and cell ID numbers are indicated at the top right of each image. Scale bars, 1 mm.

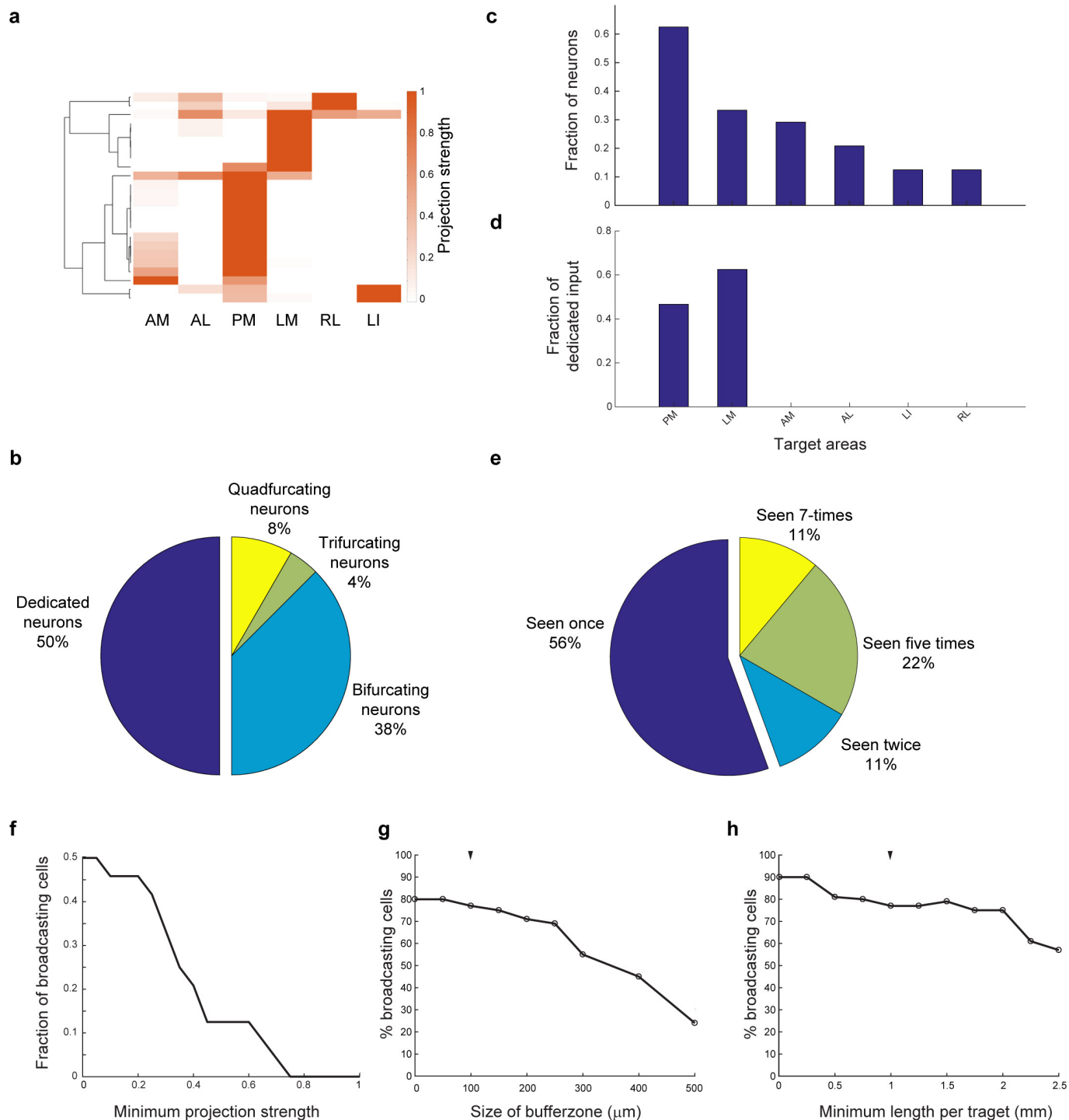
a**b**

Extended Data Figure 5 | Individual neurons in higher visual areas project to more than one target area. a, All traced neurons with cell bodies not in V1. Brain area identity is colour coded as in Fig. 1. Cell identity is indicated at the top right of each image. Scale bars, 1 mm. **b,** Histogram of the number of target areas per cell.



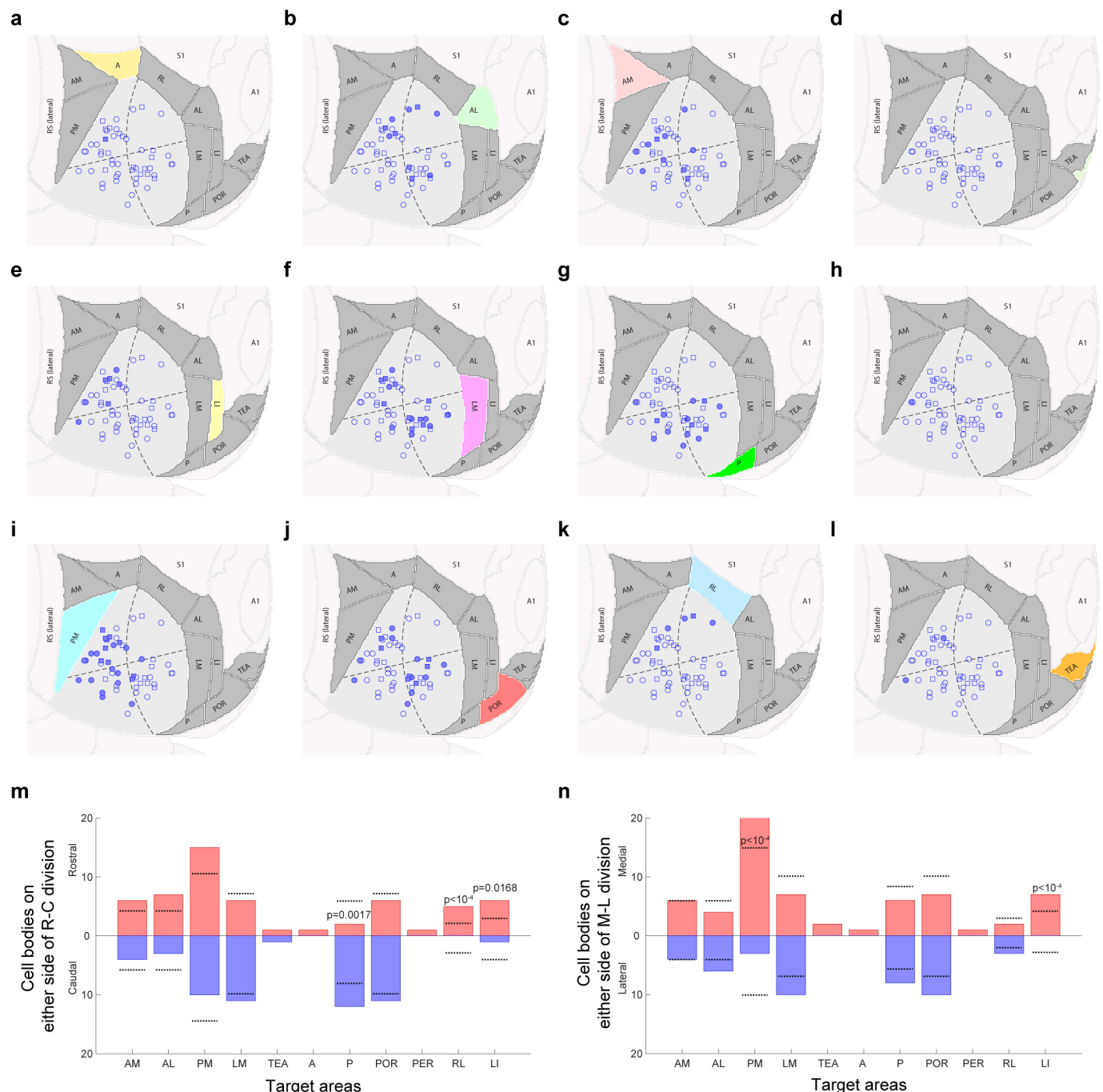
Extended Data Figure 6 | Density of axonal innervation by area and layer of V1 layer-2/3 projection neurons. **a**, Total axon length plotted as a function of the number of targets innervated by every V1 projection neuron. **b**, Axon length in area LM, PM or POR plotted as a function of the total number of targets innervated by each neuron projection to the respective area. **c–h**, The axons of V1 neurons in target areas most densely innervate layers 2/3 and 5, with some density in layer 1, but less in layers 4

and 6, often recapitulating the laminar axonal profile within V1. Coronal views of each area are shown in Allen Reference Atlas space (left) and axonal arborizations of each neuron innervating the area are colour coded. Scale bars, 200 μm . A histogram of the laminar innervation is shown (right). Note that cells with abrupt terminations outside the shown area were included in this analysis. Areas depicted are V1 (**c**), AL (**d**), LI (**e**), LM (**f**), PM (**g**) and POR (**h**). White-matter axons are not shown.



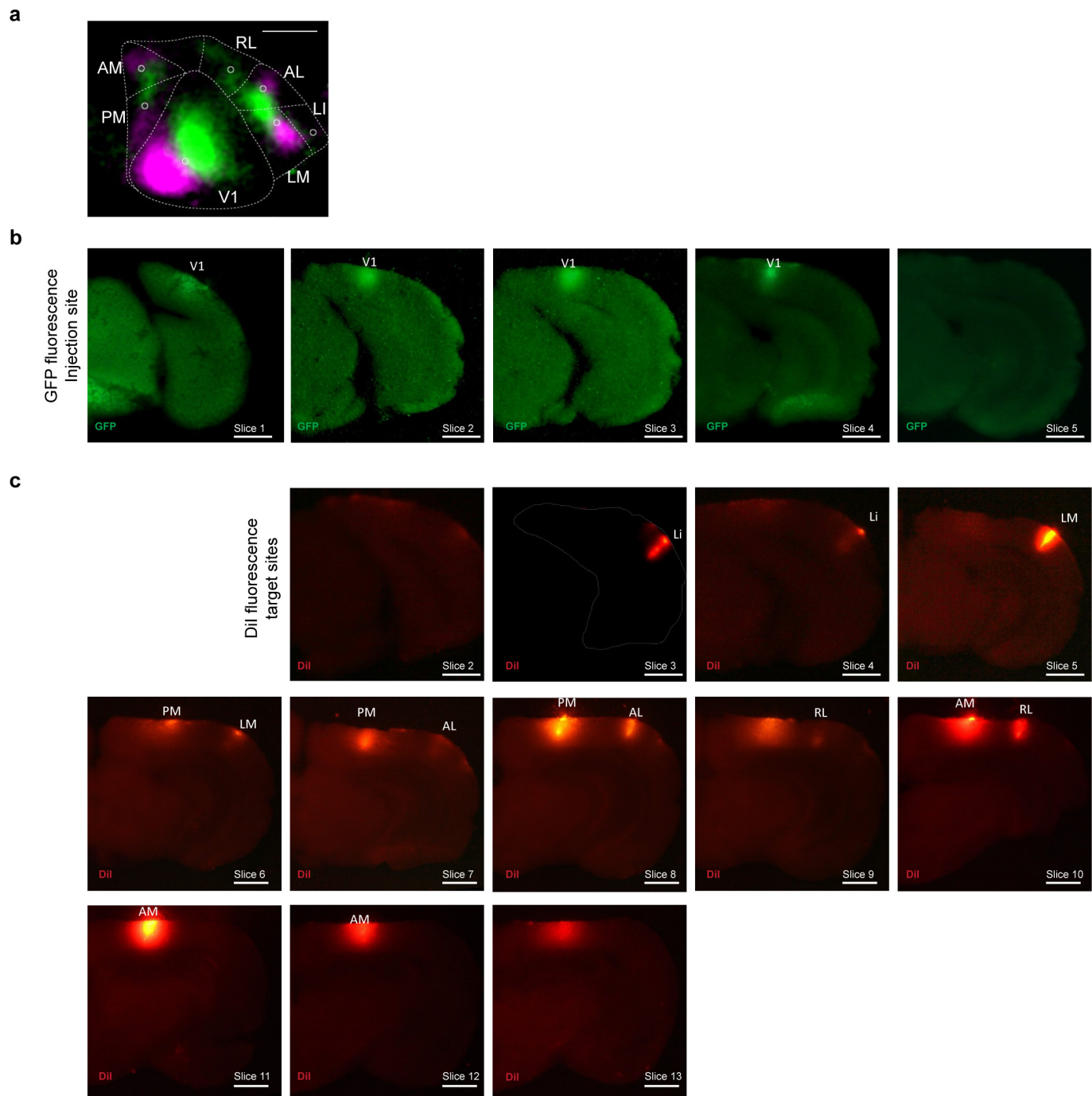
Extended Data Figure 7 | Conclusions from fluorescence-based single-neuron tracing data hold true when the analysis is restricted to subset of target areas. **a**, The projection patterns of reconstructed GFP-filled neurons when only the six target areas (LI, LM, AL, PM, AM and RL) are considered. Projection strengths are normalized to the maximum projection of each neuron, and only neurons projecting to at least one target area are shown. **b**, The distribution of target area numbers per projection neuron. **c**, The fraction of all cells projecting to each target area. **d**, The fraction of dedicated input per area. **e**, The number of times each binarized projection motif is observed. **f**, The fraction of broadcasting

cells as a function of the minimum projection strength (relative to the primary target) that each area needs to receive to be considered a target. **g**, The fraction of broadcasting cells as a function of increasing buffer zones between areas within which axons are ignored, assuming a minimum projection of 1 mm of axon per target area. **h**, The fraction of broadcasting cells as a function of the minimal amount of axon per area for it to be considered a target, assuming buffer zones of 100- μm width. Black arrowheads indicate chosen buffer zones and minimal projection for analysis in the paper.



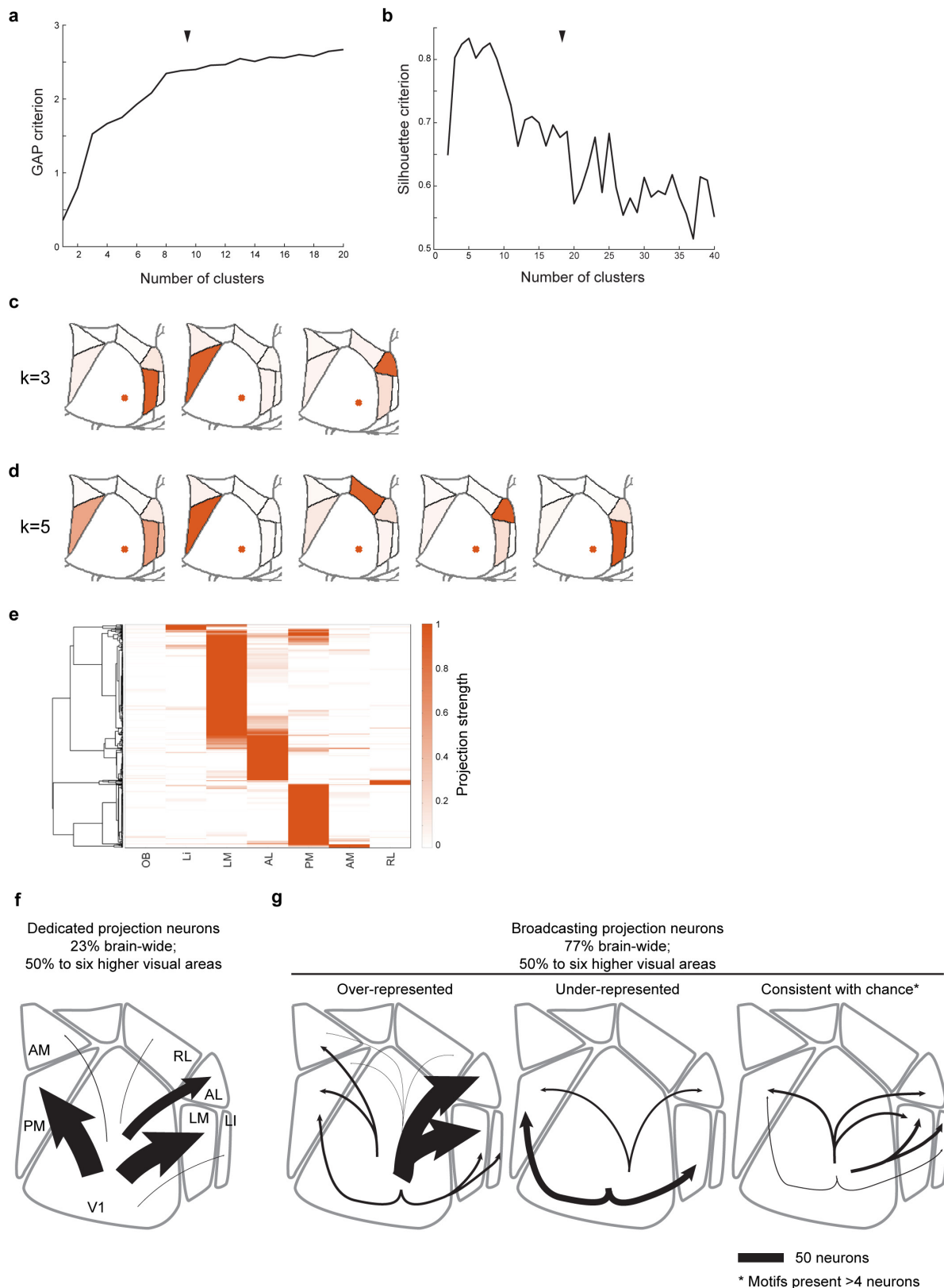
Extended Data Figure 8 | Location of cell bodies in V1 as a function of their projection targets. **a–l**, Horizontal views of Allen Reference Atlas space are shown. The location of all traced V1 neurons are indicated as circles (cells with no abrupt terminations) or squares (cells with abrupt terminations). In each plot, the cells projecting to the highlighted higher visual area are coloured in solid blue. Target areas considered are A (a), AL (b), AM (c), ECT (d), LI (e), LM (f), P (g), PER (h), PM (i), POR (j), RL (k) and TEA (l). **m, n**, Quantification of cell body location in the

rostrocaudal (**m**) and mediolateral (**n**) direction. Dotted lines indicate expected number of cells based on a bootstrapping procedure, for which we randomly selected neurons from the available positions to project to each area and repeated the process 10,000 times. P values were derived from the bootstrapping probability distribution and are indicated for projection targets significantly deviating from this expectation ($\alpha = 0.05$). P values below 10^{-4} are not exact and are therefore indicated as a range.



Extended Data Figure 9 | MAPseq dissection strategy. We identified the to-be-dissected higher visual areas by performing intrinsic imaging of visual cortex in response to stimuli at different positions in the contralateral visual field and mapping the resulting changes in intrinsic signals. **a**, A representative retinotopic map, with responses to the two 25° visual stimuli pseudocoloured in green and magenta (stimulus 1 position: 90° azimuth, 20° elevation; stimulus 2 position: 60° azimuth, 20° elevation). On the basis of this map, we fluorescently labelled retinotopically matched positions in the to-be-dissected cortical areas

with a DiI stab (white circles). Putative borders between the higher visual areas are indicated with dashed lines for orientation. Scale bar, 1 mm. $n = 4$ mice. **b**, The MAPseq virus injection site is discernible in consecutive frozen 180- μ m thick coronal sections, using GFP fluorescence. Scale bars, 1 mm. **c**, DiI injections targeted to matched retinotopic positions in six target areas identified by intrinsic signal imaging. DiI epifluorescence images of each 180- μ m thick slice are shown, and dissected areas are labelled. Scale bars, 1 mm.



Extended Data Figure 10 | Clustering of MAPseq data and data summary. **a, b**, Gap (**a**) and silhouette criteria (**b**) for *k*-means clustering of the MAPseq neurons as a function of the number of clusters. Black arrow heads indicate chosen number of clusters (*k* = 8). **c, d**, Centroids for alternative, near-optimal cluster number choices with *k* = 3 (**c**) and *k* = 5 (**d**). **e**, Hierarchical clustering results of the MAPseq dataset using a cosine distance metric. **c–e**, Colour intensity indicates projection strengths. **f, g**, Summary of single-neuron projections from V1. **f**, Cells targeting single higher visual areas (dedicated projection neurons) comprise the

minority of layer-2/3 V1 projection neurons. Among the areas analysed by MAPseq, dedicated projection neurons predominantly innervate cortical areas LM or PM. **g**, Cells projecting to two or more areas (broadcasting projection neurons) are the dominant mode of information transfer from V1 to higher visual areas. In the six areas analysed by MAPseq, broadcasting neurons innervate combinations of target areas in a non-random manner, including those that are more or less abundant than expected by chance. Line width indicates the absolute abundance of each projection type as observed in the MAPseq dataset.

Evolved Cas9 variants with broad PAM compatibility and high DNA specificity

Johnny H. Hu^{1,2,3}, Shannon M. Miller^{1,2,3}, Maarten H. Geurts^{1,2,3}, Weixin Tang^{1,2,3}, Liwei Chen^{1,2,3}, Ning Sun^{1,2,3}, Christina M. Zeina^{1,2,3}, Xue Gao^{1,2,3}, Holly A. Rees^{1,2,3}, Zhi Lin^{1,2,3} & David R. Liu^{1,2,3}

A key limitation of the use of the CRISPR–Cas9 system for genome editing and other applications is the requirement that a protospacer adjacent motif (PAM) be present at the target site. For the most commonly used Cas9 from *Streptococcus pyogenes* (SpCas9), the required PAM sequence is NGG. No natural or engineered Cas9 variants that have been shown to function efficiently in mammalian cells offer a PAM less restrictive than NGG. Here we use phage-assisted continuous evolution to evolve an expanded PAM SpCas9 variant (xCas9) that can recognize a broad range of PAM sequences including NG, GAA and GAT. The PAM compatibility of xCas9 is the broadest reported, to our knowledge, among Cas9 proteins that are active in mammalian cells, and supports applications in human cells including targeted transcriptional activation, nuclease-mediated gene disruption, and cytidine and adenine base editing. Notably, despite its broadened PAM compatibility, xCas9 has much greater DNA specificity than SpCas9, with substantially lower genome-wide off-target activity at all NGG target sites tested, as well as minimal off-target activity when targeting genomic sites with non-NGG PAMs. These findings expand the DNA targeting scope of CRISPR systems and establish that there is no necessary trade-off between Cas9 editing efficiency, PAM compatibility and DNA specificity.

The CRISPR–Cas9 system has facilitated widely used genome manipulation capabilities including targeted gene disruption^{1,2}, transcriptional activation and repression³, epigenetic modification³, and direct conversion of a target base pair to a different base pair^{4,5} in a broad range of organisms and cell types⁶. CRISPR–Cas9 targets DNA in a manner that is programmed by an RNA (typically a single-guide RNA, or sgRNA⁷) that contains a spacer sequence complementary to the target DNA site, the protospacer. In addition to a protospacer that complements the sgRNA, a Cas9 target site must also contain a PAM sequence to support recognition by Cas9. The NGG PAM requirement of canonical SpCas9, which occurs on average only once in every 16 randomly chosen genomic loci, greatly limits the targeting scope of Cas9 especially for applications that require precise Cas9 positioning, such as base editing, which requires a PAM approximately 13–17 nucleotides from the target base^{4,5}, and some forms of homology-directed repair, which are most efficient when DNA cleavage occurs roughly 10–20 base pairs away from a desired alteration^{8,9}. These requirements limit the fraction of genomic DNA that can be targeted with CRISPR systems and highlight the need for more general genome-editing tools.

To address this limitation, researchers have harnessed natural CRISPR nucleases with different PAM requirements and engineered existing systems to accept variants of naturally recognized PAMs. Other natural CRISPR nucleases shown to function efficiently in mammalian cells include *Staphylococcus aureus* Cas9 (SaCas9)¹⁰, *Acidaminococcus* sp. Cpf1¹¹, *Lachnospiraceae bacterium* Cpf1¹¹, *Campylobacter jejuni* Cas9¹², *Streptococcus thermophilus* Cas9¹³ and *Neisseria meningitidis* Cas9¹⁴. None of these CRISPR nucleases, however, offers a PAM that occurs as frequently as that of SpCas9. Although CRISPR nucleases engineered to accept additional PAM sequences^{15,16} also expand the scope of genomic targets available for Cas9-mediated manipulation, many target sequences remain inaccessible.

Here we used phage-assisted continuous evolution (PACE) to rapidly generate Cas9 variants that accept an expanded range of PAM

sequences. During PACE, host *Escherichia coli* cells continuously dilute an evolving population of bacteriophages (selection phages). Because dilution occurs faster than cell division but slower than phage replication, only the selection phages, and not the host cells, can accumulate mutations¹⁷. Each selection phage carries a gene to be evolved instead of a phage gene (gene III) that is required for the production of infectious progeny phage. Selection phages that contain desired gene variants trigger host-cell gene III expression from the accessory plasmid and the production of infectious selection phages that propagate the desired variants. Phages encoding inactive variants do not generate infectious progeny and are rapidly diluted out of the culture vessel (Fig. 1a). As phage replication can occur in as little as 10 min, PACE enables hundreds of generations of directed evolution to occur per week without researcher intervention^{17–22}.

Evolution of Cas9 towards expanded PAM compatibility

To link Cas9 DNA recognition to phage propagation during PACE, we developed a bacterial one-hybrid selection^{20,23,24} in which the selection phage encodes a catalytically dead SpCas9 (dCas9) fused to the ω subunit of bacterial RNA polymerase. When this fusion binds an sgRNA encoded by the accessory plasmid and a PAM and protospacer upstream of gene III in the accessory plasmid, recruitment RNA polymerase causes gene III expression and phage propagation (Fig. 1b). We envisioned installing a library of all 64 possible NNN PAM sequences at the target protospacer in the accessory plasmid, so that selection phages encoding Cas9 variants with broader PAM compatibility would replicate in a larger fraction of host cells and thus experience a fitness advantage.

We optimized the relationship between Cas9 DNA binding and gene expression (Extended Data Fig. 1a–c). These studies revealed that a fusion of the orientation N– ω –dCas9–C, together with a simple Ala–Ala fusion linker and the placement of the protospacer on the reverse complement strand 45 base pairs upstream of the –35 box results in the strongest guide RNA-dependent gene expression activation (13-fold)

¹Merkin Institute of Transformative Technologies in Healthcare, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ²Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA. ³Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

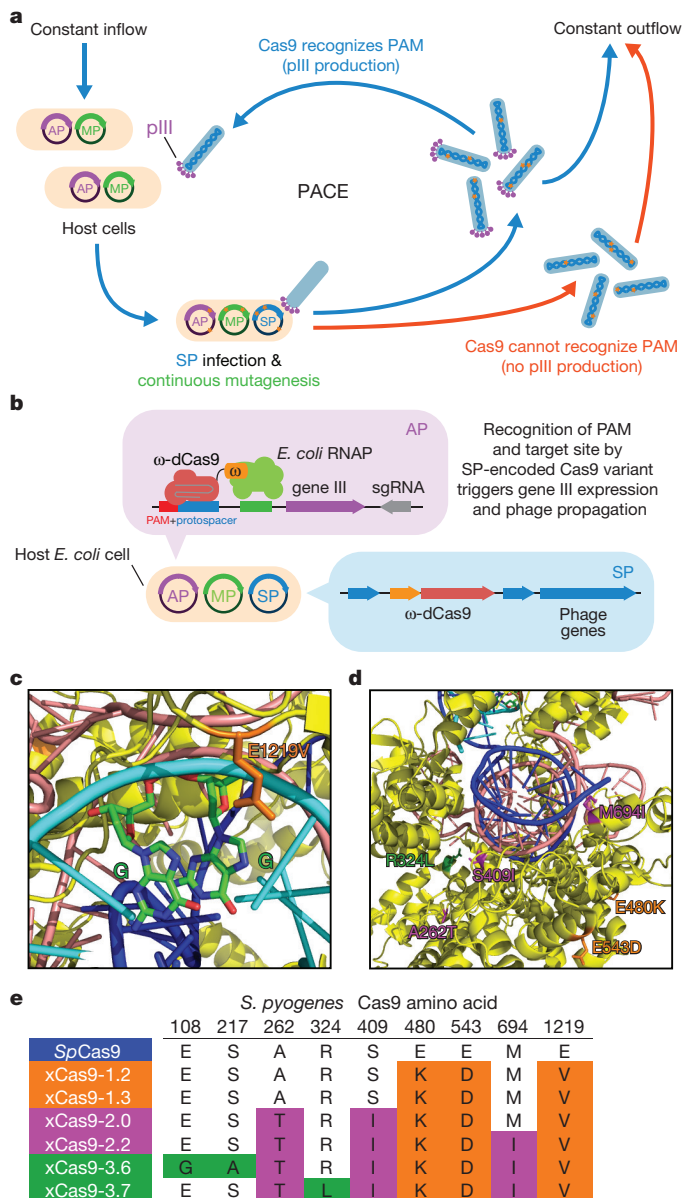


Figure 1 | PACE of Cas9 variants with broadened PAM compatibility.
a, PACE takes place in a fixed-volume 'lagoon' that is continuously diluted with fresh host *E. coli* cells. Upon infection, selection phages (SP) that encode a Cas9 variant capable of binding the target PAM and protospacer on the accessory plasmid (AP) induce expression of gene III, resulting in infectious progeny phages that propagate the active Cas9 variant in subsequent host cells. pIII, attachment protein G3P (encoded by gene III).
b, Representation of a phage-infected host cell during PACE. The host cell carries the accessory plasmid, which links Cas9 target DNA binding to phage propagation, and the mutagenesis plasmid (MP), which increases mutagenesis during PACE. RNAP, RNA polymerase. **c**, **d**, The crystal structure of SpCas9 with the location of xCas9 mutations shown.
e, Genotypes of some evolved xCas9 variants, coloured by evolution stage. See Supplementary Table 5 for the 95 xCas9 variant genotypes.

(Extended Data Fig. 1a–c). Together, these results establish a linkage between Cas9 DNA-binding activity and gene expression in a selection system suitable for PACE.

Using this selection, we first allowed a selection phage encoding the ω -dCas9 fusion to self-optimize on host cells containing an accessory plasmid with a canonical NGG PAM, resulting in enrichment of an I12N mutation in the ω subunit. Adding this single mutation to ω -dCas9 boosted activation from 13-fold to over 100-fold (Extended

Data Fig. 1d), representing early-stage optimization of ω -dCas9 during PACE before evolution for broadened PAM compatibility.

To evolve Cas9 variants with expanded PAM compatibility we generated three accessory plasmid libraries, each containing a different sgRNA (Supplementary Table 4) and a corresponding protospacer upstream of an NNN PAM library, in which N is an equimolar mixture of all four DNA bases. This design imposes selection pressure to recognize many different PAM sequences, as well as to maintain compatibility with different target DNA sequences. All three accessory plasmid libraries were introduced into host *E. coli* cells harbouring the mutagenesis plasmid MP6¹⁷. The resulting host cells were incubated overnight with selection phages containing ω (I12N)-dCas9. This phage-assisted non-continuous evolution system^{19,21} preferentially replicates Cas9 variants that bind a greater variety of PAM sequences, similar to PACE, but with lower stringency since there is no outflow of the phages.

After 24 days of serial overnight propagation and 1:1,000 dilution, we isolated five Cas9 clones for sequencing and characterization (xCas9-1.0–xCas9-1.4). Notable recurring mutations include E480K, E543D and E1219V (Fig. 1e and Supplementary Table 5). E1219 is close in the SpCas9 crystal structure to R1333 and R1335 (Fig. 1c), two residues that are known to have a critical role in PAM recognition²⁵. The mixture of phage from the final phage-assisted non-continuous evolution pool were further evolved for 72 h using PACE on host cells containing the same accessory plasmid libraries harbouring NNN PAM sequences. Among the individual Cas9 clones emerging from PACE (xCas9-2.0–xCas9-2.6), E480K, E543D and E1219V were present in all sequenced phages, along with additional mutations seen in multiple clones, including A262T, K294R, S409I and M694I (Fig. 1e and Supplementary Table 5). Finally, the resulting phages were continuously evolved using PACE for an additional 72 h on host cells containing three protospacer–sgRNA pairs and HHH PAM libraries, in which H is A, C, or T, to favour Cas9 variants with activity on non-NGG PAMs.

Fourteen resulting evolved Cas9 variants (xCas9-3.0–xCas9-3.13) containing consensus mutations (Fig. 1e and Supplementary Table 5) emerged from two apparent evolution trajectories. Although all phages shared E480K, E543D and E1219V core mutations, xCas9-3.0–xCas9-3.5 also all contained K294R and Q1256K mutations and xCas9-3.6–xCas9-3.13 all contained A262T, S409I and M694I mutations, with some of the latter also containing R324L. The Cas9 crystal structure predicts that R324L, S409I and M694I lie near the DNA–sgRNA interface (Fig. 1d) and could have a role in mediating DNA sequence recognition and the switching of Cas9 from the open to the closed conformation upon target recognition^{26,27}. Because the entire Cas9 gene was subject to mutagenesis during PACE, the mechanism of xCas9 may differ from that of engineered Cas9 variants that primarily mutated DNA-contacting residues^{15,16,28}.

We characterized evolved xCas9 variants in several contexts. We first restored the catalytic residues D10 and H840 to test whether xCas9 nucleases can cleave DNA even though they were evolved only for DNA binding. The xCas9-3.0–xCas9-3.13 clones were tested in a PAM depletion assay^{15,16} in which they were given the opportunity to cleave a library of plasmids containing a protospacer and all possible NNN PAM sequences in an antibiotic-resistance gene in bacterial cells. Plasmid cleavage results in the loss of spectinomycin resistance. This PAM depletion assay revealed that xCas9-3.0–xCas9-3.3 and xCas9-3.5–xCas9-3.9 cleave DNA site with NG, NNG, GAA, GAT and CAA PAMs (Extended Data Fig. 2). The clone with the highest PAM depletion score, xCas9-3.7, depleted NG, NNG, GAA, GAT and CAA PAMs by at least 100-fold compared to the starting library, and xCas9-3.6 showed the second highest average PAM depletion score.

xCas9 activators and nucleases in human cells

To test whether mutations evolved during PACE in bacteria are compatible with xCas9 function in mammalian cells, we characterized xCas9 variants for their activity and PAM compatibility in human cells in

four contexts: transcriptional activation, genomic DNA cutting, cytidine base editing and adenine base editing. All guide RNAs used in this study were transcribed from a U6 promoter, and natively started with a G at the 5' end to avoid possible losses in activity caused by a mismatched 5' guide terminus^{15,16,29,30}. To test for transcriptional activation, catalytically dead versions of xCas9 were fused to the transcriptional activator VP64–p65–Rta (dxCas9–VPR)³¹. Plasmids encoding dxCas9–VPR, a green fluorescent protein (GFP) reporter downstream of a target protospacer and a corresponding sgRNA were co-transfected into human HEK293T cells³¹. Target-gene transcriptional activation was measured by cellular GFP fluorescence after three days. Three different target-site PAM sets were tested: a single reporter with an NGG PAM, a reporter library containing a NNN PAM library and a reporter library containing a NNNNN PAM library. In addition, two different protospacer sequences, reporter 1 and reporter 2, were tested with their corresponding sgRNAs.

Most early stage xCas9 variants outperformed wild-type *SpCas9* on sites with NGG PAMs, as well as with NNN and NNNNN PAM libraries (Extended Data Fig. 3a). For reporters 1 and 2, respectively, xCas9-3.7 achieved 2.8- and 1.5-fold higher mean fluorescence for the NGG PAM, 7.9- and 2.1-fold higher mean fluorescence for the NNN PAM library, and 5.2- and 1.7-fold higher mean fluorescence for the NNNNN PAM library when compared with *SpCas9* (Extended Data Fig. 3b, c). The similar performance of xCas9-3.7 on NNN and NNNNN PAM libraries suggests that it did not evolve strong sequence preferences at nucleotides immediately downstream of the NNN PAM. The xCas9-3.6 variant showed similar results to those of xCas9-3.7 in this assay (Extended Data Fig. 3b, c).

To dissect activity on individual PAM sequences, we tested dxCas9–VPR transcriptional activators on individual target sites containing each of the 64 possible three-nucleotide PAM sequences (Fig. 2a and Extended Data Figs 4–6) in HEK293T cells. Consistent with the PAM library results, dxCas9(3.7)–VPR showed broad improvements in transcriptional activity relative to d*SpCas9*–VPR across many individual non-NGG PAMs. Transcriptional activation by dxCas9(3.7)–VPR at sites containing NGT, NGA, NGC, NNG, GAA and GAT PAMs averaged 56–91% of the average activity of dxCas9(3.7)–VPR on the four NGG PAM sites (Fig. 2a and Extended Data Fig. 5). The performance of xCas9-3.6 transcriptional activators was similar to that of xCas9-3.7 (Extended Data Fig. 6). To test whether broadened transcriptional activation by xCas9 is limited to reporter plasmids that may be only partially chromatinized, we also tested the ability of dxCas9(3.7)–VPR to activate transcription of six endogenous genomic loci in human cells and observed 3.3-fold average improved activation of the two NGG PAM sites and 39-fold average improved activation among the three NGN PAM sites, but no improvement on the tested NNG site, relative to d*SpCas9*–VPR (Extended Data Fig. 5e). Overall, these results establish that xCas9 is compatible with the dCas9–VPR architecture and can serve as a potent transcriptional activator in human cells at a substantially expanded set of PAMs. On the basis of their strong performance in the PAM depletion assay and as transcriptional activators, we chose xCas9-3.7 and xCas9-3.6 for further characterization.

To test targeted genomic DNA cleavage in human cells, we expressed xCas9-3.7 and -3.6 nuclease in a HEK293T cell line with a genomically integrated GFP gene and measured the loss of GFP fluorescence reflecting DNA cleavage and indel-mediated disruption of the target site. There are two NGG PAM sites and three NGT sites present in the *GFP* sequence, as well as individual NGC, NGA, GAT, NCG and NTG sites. All of these PAM sequences were tested with *SpCas9*, xCas9-3.7 and xCas9-3.6. For the NGG PAM sites, xCas9-3.7 modestly outperformed *SpCas9*, resulting in $46 \pm 2.0\%$ compared to $33 \pm 3.4\%$ GFP disruption for *SpCas9* (mean \pm s.d. of three independent replicates, Fig. 2b). For all tested non-NGG PAM sites, xCas9-3.7 showed substantially higher (1.6- to 5.1-fold) average apparent cleavage activity than *SpCas9* (Fig. 2b). At all tested sites, xCas9-3.6 showed GFP disruption percentages that were similar to or slightly lower than

those of xCas9-3.7 (Extended Data Fig. 7a). Neither xCas9-3.7 nor xCas9-3.6 increased GFP loss relative to *SpCas9* for either NNG PAM site tested, suggesting that the transcriptional activation by dxCas9–VPR observed at some NNG PAM sites, and the strong NNG PAM signal in the bacterial PAM depletion assay, do not necessarily translate to DNA cleavage at all NNG sites in mammalian cells. Target-site dependence is also well-known for *SpCas9*³².

To further characterize DNA cleavage by xCas9 variants in human cells, we targeted endogenous genomic sites in HEK293T cells and measured indel formation by high-throughput sequencing (HTS). Twenty endogenous sites were tested covering four NGG PAM sites, all twelve possible NGT, NGC and NGA PAMs, and GAT, GAA and CAA PAMs (Fig. 2c). On the four NGG PAM sites tested, xCas9-3.7 showed comparable activity to *SpCas9*, averaging $41 \pm 6.4\%$ indels compared to $41 \pm 6.0\%$ for *SpCas9* (Fig. 2c). All four NGT PAM sites showed much higher indel formation with xCas9-3.7 than with *SpCas9*, averaging $38 \pm 4.1\%$ indels compared to $8.6 \pm 1.5\%$ for *SpCas9*, a 4.5-fold increase. The four NGA PAM sites averaged $32 \pm 2.4\%$ indels for xCas9-3.7 and $20 \pm 2.7\%$ for *SpCas9*, a 1.6-fold increase, consistent with previous reports that NGA can serve as a secondary PAM for *SpCas9*³³. Although indel frequencies at the endogenous NGC PAM sites were more variable, ranging from 4.8 to 31%, xCas9-3.7 averaged $13 \pm 0.90\%$ indels compared to $6.3 \pm 0.77\%$ for *SpCas9*, a 2.1-fold increase. Among the three GAA and GAT sites tested, *SpCas9* showed virtually no activity, averaging $1.4 \pm 1.3\%$ indel formation, whereas xCas9-3.7 averaged $7.2 \pm 2.8\%$ indel formation, a 5.2-fold increase. For all sites tested the xCas9-3.6 variant showed indel frequencies that were similar to or slightly lower than those of xCas9-3.7 (Extended Data Fig. 7b). Negative control experiments lacking sgRNA plasmids resulted in no indels above background (Extended Data Fig. 8). Taken together, these results indicate that xCas9-3.7 nuclease mediates target-gene disruption at NGG PAM sites with efficiencies that are comparable to wild-type *SpCas9*, but cleaves NG, GAA and GAT PAM sites with substantially higher efficiencies than *SpCas9*. The greater PAM-dependent and protospacer-dependent variability of xCas9 nuclease-mediated gene disruption relative to transcriptional activation (Fig. 2 and Extended Data Figs 5–7) may reflect more extensive requirements for DNA cleavage than DNA binding^{27,29}, or differences in the chromatin state of plasmid (Fig. 2a and Extended Data Figs 4, 5a–d, 6) versus genomic targets (Fig. 2b, c and Extended Data Figs 5e, 7).

Base editing by xCas9 variants in human cells

Base editing is a newer genome editing approach that uses a catalytically impaired Cas9 fused to a natural or laboratory-evolved nucleobase deaminase enzyme and, in some cases, a DNA glycosylase inhibitor to directly convert a target C•G to T•A, or a target A•T to G•C, without introducing double-stranded DNA breaks or requiring homology-directed repair^{4,5,34}. The suitability of a target site for base editing is highly dependent on the presence of a suitably positioned PAM, which must exist within a narrow window downstream of the target base pair (typically 15 ± 2 nucleotides). The broad PAM compatibility of xCas9 variants thus has the potential to expand the DNA targeting scope of base editors (Fig. 3c).

To evaluate C•G-to-T•A base-editing activity of xCas9 variants, we substituted *SpCas9* with xCas9-3.7 and -3.6 in the third-generation (BE3) base editor architecture⁴. We separately transfected plasmids encoding xCas9(3.7)–BE3, xCas9(3.6)–BE3 and *SpCas9*–BE3 into mammalian cells to compare editing efficiency on all 20 sites tested above for endogenous genomic DNA cleavage (Fig. 3a). At the four NGG PAM target sites tested, xCas9(3.7)–BE3 averaged $37 \pm 10\%$ C•G-to-T•A conversion, whereas *SpCas9*–BE3 averaged $28 \pm 5.2\%$ (Fig. 3a). At the NG, GAA and GAT PAM sites tested, xCas9(3.7)–BE3 resulted in substantially improved base editing, averaging $24 \pm 5.4\%$ editing at NGT PAM sites, an 9.5-fold increase compared with that of *SpCas9*; $16 \pm 3.5\%$ editing at NGA sites, a 3.5-fold increase compared

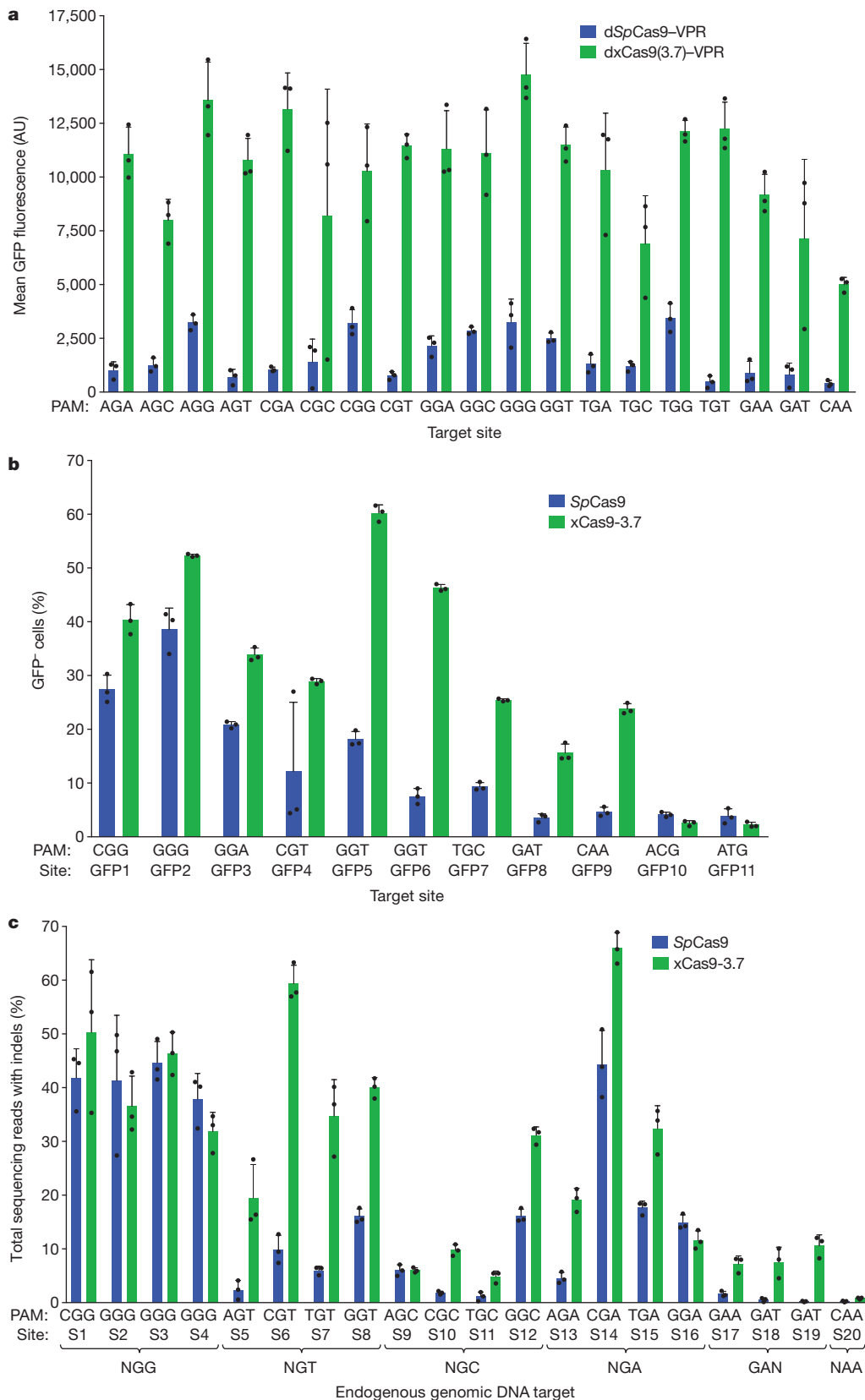


Figure 2 | Transcriptional activation and genomic DNA cleavage by evolved xCas9-3.7 in human cells. **a**, Transcriptional activation by dSpCas9-VPR and dxCas9(3.7)-VPR targeting GFP reporter plasmids containing the same protospacer but different PAM sites in HEK293T cells. AU, arbitrary units. **b**, Genomic DNA cleavage by SpCas9 or xCas9-3.7 in HEK293-GFP cells that have a genomically integrated GFP reporter gene. After 5 days, the cells were analysed for loss of GFP fluorescence by flow cytometry. **c**, DNA cleavage of endogenous genomic DNA sites with NGG and non-NGG PAMs by SpCas9 and xCas9-3.7 in HEK293T cells. Indel rates were measured by HTS 5 days after plasmid transfection. **a–c**, Data are mean and s.d. of three biologically independent samples. Target sites are listed in Supplementary Tables 8, 11 and 12.

with SpCas9; $6.2 \pm 0.34\%$ editing at NGC sites, a 13-fold increase over SpCas9; $10 \pm 0.75\%$ editing on the GAA PAM site, a more than 50-fold increase over SpCas9; and $12 \pm 1.5\%$ editing on the GAT sites, a greater than 100-fold increase compared with SpCas9 (Fig. 3a). The base editing efficiencies of xCas9(3.6)-BE3 were comparable to, or slightly worse than, those of xCas9(3.7)-BE3 (Extended Data Fig. 7c). We also tested

cytosine base editing at an additional 15 endogenous genomic sites within the *FANCF* gene and observed similar large improvements for xCas9(3.7)-BE3 over SpCas9-BE3 (Extended Data Fig. 9a). Overall, these results indicate that xCas9 variants are compatible with the BE3 architecture enabling cytidine base editing of target sites that cannot be accessed by SpCas9-BE3 or, with the exception of NGA PAM

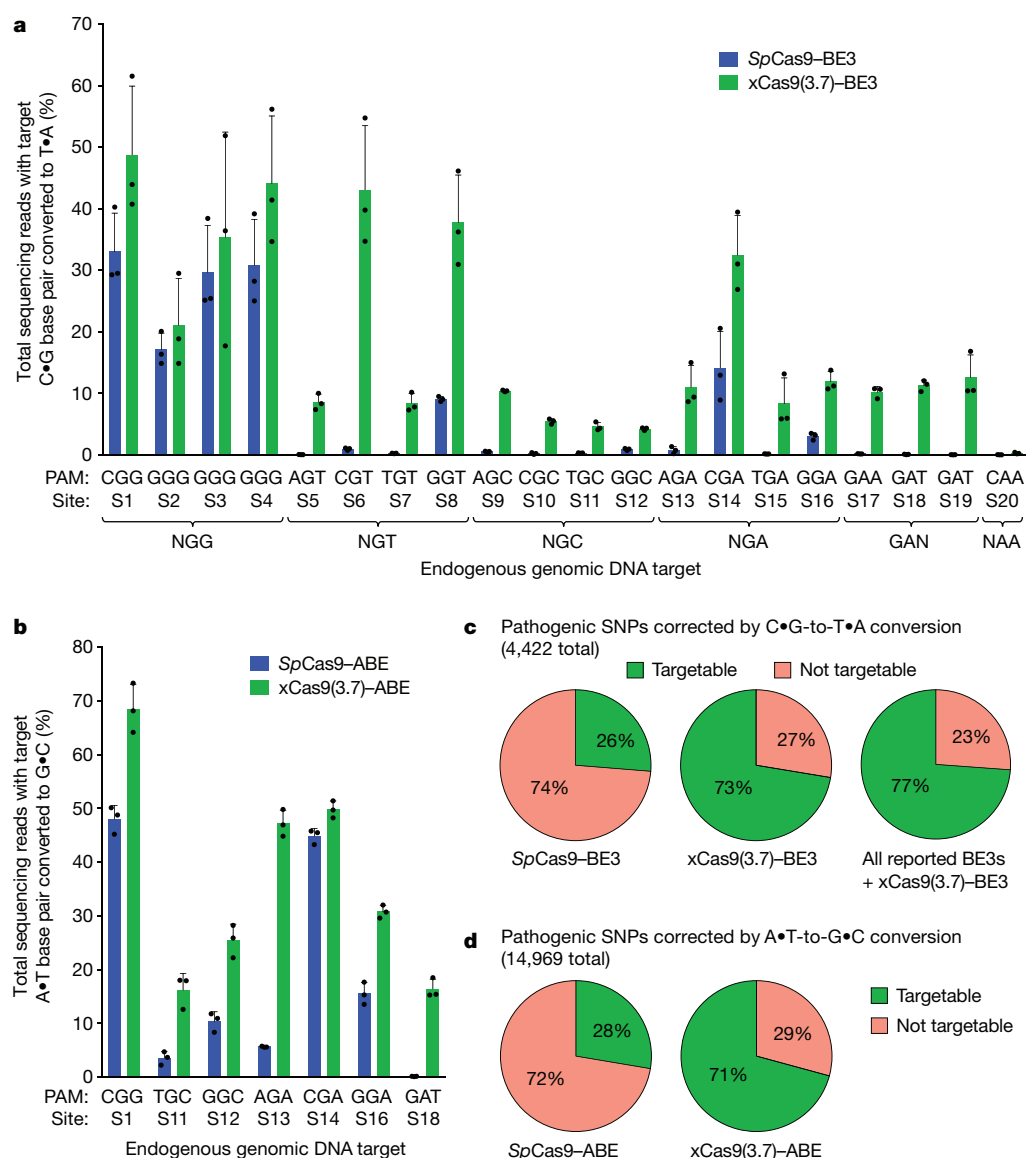


Figure 3 | Cytidine and adenine base editing by xCas9. **a**, C•G-to-T•A conversion frequencies at the most efficiently edited base for 20 endogenous genomic loci in HEK293T cells 3 days after plasmid transfection. **b**, A•T-to-G•C conversion frequencies at the most efficiently edited base for seven endogenous genomic loci in HEK293T cells 5 days after plasmid transfection. **a**, **b**, Data are mean and s.d. of three biologically independent samples. See Supplementary Table 14 for complete HTS results across the protospacer. **c**, Fraction of pathogenic T•A-to-C•G to SNPs in ClinVar³⁹ that could, in principle, be corrected by *SpCas9*-BE3 (left), *xCas9*(3.7)-BE3 (middle) or *xCas9*(3.7)-BE3 with all BE3 variants reported to date (right). **d**, Fraction of G•C-to-A•T pathogenic SNPs in ClinVar that could, in principle, be corrected by *SpCas9*-ABE (left) or *xCas9*(3.7)-ABE (right).

sites³⁴, by any other previously reported base editors. We also tested *xCas9*-3.7 in the BE4 architecture that was designed to reduce undesired byproducts³⁵. In this system, fewer indels and higher product purities were observed, although editing efficiencies were slightly lower (Extended Data Fig. 9b–d).

The recent development of an adenine base editor (ABE) enables programmable installation of A•T-to-G•C mutations⁵. No ABEs have been reported that can target non-NGG PAM sites, limiting its targeting scope. We replaced *SpCas9* in ABE 7.10 (ref. 5) with *xCas9*-3.7 and -3.6, and assayed the resulting *xCas9*-ABEs in HEK293T cells at the seven endogenous genomic sites tested above that contain an A in the targeting window of ABE (positions 4–8, counting the PAM as positions 21–23). At all seven of these sites, *xCas9*(3.7)-ABE resulted in higher base-editing efficiencies than the original *SpCas9*-ABE (Fig. 3b). Average base-editing efficiency at the NGG PAM site tested increased from $48 \pm 2.1\%$ to $69 \pm 3.7\%$. At the GAT PAM site tested, *xCas9*(3.7)-ABE resulted in $16 \pm 1.5\%$ base editing, whereas *SpCas9*-ABE yielded no detectable editing (at most 0.1%), representing a more than 100-fold increase. On the two NGC and three NGA sites tested, *xCas9*(3.7)-ABE averaged $21 \pm 2.5\%$ and $43 \pm 1.5\%$ base editing, respectively, whereas *SpCas9*-ABE averaged $7.0 \pm 1.3\%$ and $22 \pm 1.2\%$, respectively. Base editing by *xCas9*(3.7)-ABE was comparable to or higher than that of *xCas9*(3.6)-ABE (Extended Data Fig. 7d). Collectively, these results

establish that *xCas9*-ABE mediates adenine base editing at sites that cannot currently be accessed.

Improved DNA specificity of *xCas9* in human cells

Because PAM recognition is a crucial component of Cas9 DNA specificity²⁷, the substantially broadened PAM compatibility of *xCas9* proteins would be expected to increase their off-target activity^{29,36}. Indeed, the engineered *S. aureus* KKH-*SaCas9*, which accepts an NNNRRT PAM instead of the native NNGRRT *SaCas9* PAM, exhibits similar or higher levels of off-target editing than *SaCas9*¹⁵. Most of the *xCas9* mutations are close to the PAM or to the DNA-sgRNA interface (Fig. 1), raising the possibility that these mutations might alter the degree to which mismatches between the spacer and protospacer impede productive DNA binding or editing. Previous studies have demonstrated that some mutations near the Cas9-DNA interface can reduce off-target activity, in some cases without affecting on-target modification efficiency^{26,37,38}.

To test the off-target activity of *xCas9* variants, we performed genome-wide, unbiased identification of double-strand breaks enabled by sequencing (GUIDE-seq), an off-target analysis method²⁹, of *xCas9*-3.7, *xCas9*-3.6 and *SpCas9* in HEK293T and U2OS cells. Notably, for all five endogenous genomic NGG PAM sites tested in HEK293T cells and for both NGG PAM sites tested in U2OS cells,

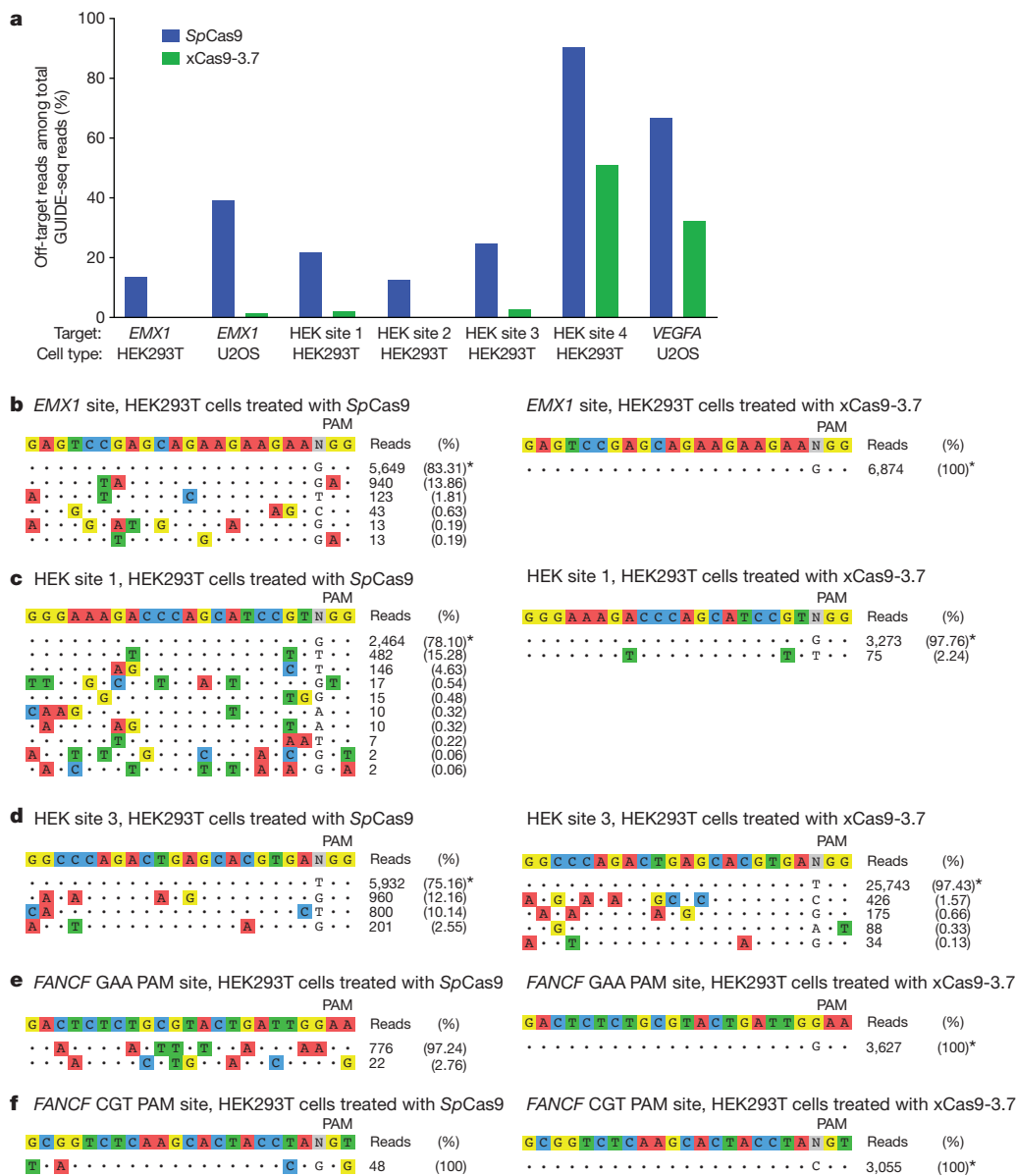


Figure 4 | Off-target editing analysis of xCas9. a, GUIDE-seq²⁹ was performed on *SpCas9* and *xCas9-3.7* nucleases. Six endogenous genomic sites with NGG PAMs were tested in HEK293T or U2OS cells. The percentage of off-target sequencing reads relative to total reads are shown. **b–d**, All GUIDE-seq on-target reads (indicated by an asterisk) and off-target reads for three sites in HEK293T cells are shown for *SpCas9* and *xCas9-3.7*. See Extended Data Fig. 10 for additional GUIDE-seq results. **e, f**, GUIDE-seq results for two endogenous genomic non-NGG PAM sites in HEK293T cells. No on-target GUIDE-seq reads were detected for *SpCas9* at either of these non-NGG sites. See Extended Data Fig. 10 for GUIDE-seq analysis of *xCas9-3.6* and Extended Data Fig. 11 for HTS validation of GUIDE-seq results. Target sequences are listed in Supplementary Table 15.

GUIDE-seq analysis revealed that *xCas9-3.7* and *-3.6* resulted in much lower off-target activity than *SpCas9*, as reflected by both the number of detected off-target sites as well as the total modification frequency at each detected off-target site (Fig. 4 and Extended Data Fig. 10). For example, at the *EMX1* target site in HEK293T cells, *SpCas9* showed 5,649 on-target reads and a total of 1,132 off-target reads, whereas *xCas9 3.7* showed 6,874 on-target reads but no off-target reads (Fig. 4b). In U2OS cells, the off-target to on-target ratio for the same site was 0.65 for *SpCas9* with 6,328 on-target reads, and 0.015 for *xCas9-3.7* with 22,539 on-target reads, representing a 43-fold reduction in off-target modification (Extended Data Fig. 10c). Likewise, for HEK sites 1, 2 and 3, *xCas9-3.7* resulted in at least 100-fold lower off-target to on-target modification ratios (0.023, less than 0.001 and 0.028, respectively) than those of *SpCas9* (0.28, 0.14 and 0.33, respectively) (Fig. 4 and Extended Data Fig. 10). For the known highly promiscuous target sites HEK site 4 and *VEGFA*²⁹, the off-target to on-target ratios were 9.4 and 2.0, respectively, for *SpCas9*, but only 1.0 and 0.48 for *xCas9-3.7*, a 4.2- to 9.4-fold improvement (Extended Data Fig. 10). We observed these large improvements in DNA specificity for *xCas9-3.7* and *-3.6* even though *xCas9* variants showed a much broader range of PAM sequences among detected off-target sites (Fig. 4 and Extended Data Fig. 10). The GUIDE-seq

results were verified by HTS of many individual on-target and off-target sites from the genomic DNA of treated cells (Extended Data Fig. 11).

We also evaluated the off-target DNA specificity of both *SpCas9* and *xCas9-3.7* at two non-NGG PAM (GAA and CGT) sites in HEK293T cells. As expected, GUIDE-seq did not yield any on-target reads for *SpCas9* at either of these non-NGG PAM sites, whereas *xCas9-3.7* had 3,627 on-target reads for the GAA PAM site and 3,055 on-target reads for the CGT PAM site (Fig. 4e, f). Notably, neither site exhibited any detected off-target GUIDE-seq reads for *xCas9-3.7*, although potential off-target reads were detected for *SpCas9* (Fig. 4e, f). Collectively, these findings reveal that *xCas9-3.7* and *-3.6* offer greatly reduced off-target activity compared with wild-type *SpCas9*, despite their broader PAM compatibility. These results also establish that there is no trade-off between Cas9-mediated editing efficiency, PAM compatibility and DNA specificity, a key finding as natural and engineered genome-editing agents advance into widespread applications including human clinical trials.

These results, together with the success of multiple independent efforts to create high-fidelity Cas9 variants^{26,37,38}, suggest that the DNA promiscuity of wild-type Cas9, which probably evolved to impede viral evasion, can be overcome by protein engineering or evolution. That

xCas9 exhibits much higher DNA specificity than SpCas9 even though it was not explicitly selected for this property suggests that the off-target activity of wild-type SpCas9 may lie at a narrow fitness peak that is suitable for defending the much smaller bacterial genome but not optimal for genome editing in mammalian cells. These observations are therefore consistent with a model in which native SpCas9 is poised to become more specific, rather than less specific, upon sufficient mutation.

To our knowledge, the targeting scope of xCas9 is the broadest among Cas9 variants known to function efficiently in mammalian cells. Evolved xCas9 variants are also the first to offer improvements in targeting scope, editing efficiency and DNA specificity in a single entity relative to wild-type SpCas9. Although the efficacy of xCas9 on non-NGG PAMs varies based on application (transcriptional activation, DNA cleavage or base editing) and on target site, the ability to access some NG, GAA and GAT PAM sequences greatly expands the breadth of targets available for site-sensitive genome editing applications. Compared to SpCas9-BE3, xCas9(3.7)-BE3 increases the percentage of the 4,422 pathogenic single-nucleotide polymorphisms (SNPs) in the ClinVar database³⁹ that could, in principle, be targeted by C•G-to-T•A base editing from 26% to 73% (Fig. 3c). Likewise, xCas9(3.7)-ABE increases the fraction of the 14,969 pathogenic SNPs in ClinVar that could be targeted by A•T-to-G•C base editing from 28% to 71% (Fig. 3d). We anticipate that xCas9 and additional CRISPR enzyme variants with broadened PAM compatibilities may also expand the scope of other forms of nucleic acid editing, CRISPR-based screens and epigenetic modification.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 January; accepted 21 February 2018.

Published online 28 February 2018.

1. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
2. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
3. Mitsuonobu, H., Teramoto, J., Nishida, K. & Kondo, A. Beyond native Cas9: manipulating genomic information and function. *Trends Biotechnol.* **35**, 983–996 (2017).
4. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
5. Gaudelli, N. M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
6. Komor, A. C., Badran, A. H. & Liu, D. R. CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* **168**, 20–36 (2017).
7. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
8. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).
9. Yang, L. et al. Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* **41**, 9049–9061 (2013).
10. Ran, F. A. et al. In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
11. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).
12. Kim, E. et al. In vivo genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nat. Commun.* **8**, 14500 (2017).
13. Müller, M. et al. *Streptococcus thermophilus* CRISPR-Cas9 systems enable specific editing of the human genome. *Mol. Ther.* **24**, 636–644 (2016).
14. Lee, C. M., Cradick, T. J. & Bao, G. The *Neisseria meningitidis* CRISPR-Cas9 system enables specific genome editing in mammalian cells. *Mol. Ther.* **24**, 645–654 (2016).
15. Kleinstiver, B. P. et al. Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).
16. Kleinstiver, B. P. et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).

17. Badran, A. H. & Liu, D. R. Development of potent *in vivo* mutagenesis plasmids with broad mutational spectra. *Nat. Commun.* **6**, 8425 (2015).
18. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).
19. Badran, A. H. et al. Continuous evolution of *Bacillus thuringiensis* toxins overcomes insect resistance. *Nature* **533**, 58–63 (2016).
20. Hubbard, B. P. et al. Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nat. Methods* **12**, 939–942 (2015).
21. Bryson, D. I. et al. Continuous directed evolution of aminoacyl-tRNA synthetases. *Nat. Chem. Biol.* **13**, 1253–1260 (2017).
22. Packer, M. S., Rees, H. A. & Liu, D. R. Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nat. Commun.* **8**, 956 (2017).
23. Meng, X. & Wolfe, S. A. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat. Protocols* **1**, 30–45 (2006).
24. Dove, S. L., Joung, J. K. & Hochschild, A. Activation of prokaryotic transcription through arbitrary protein-protein contacts. *Nature* **386**, 627–630 (1997).
25. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
26. Chen, J. S. et al. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
27. Sternberg, S. H., LaFrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110–113 (2015).
28. Gao, L. et al. Engineered Cpf1 variants with altered PAM specificities. *Nat. Biotechnol.* **35**, 789–792 (2017).
29. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
30. Kim, S. et al. Rescue of high-specificity Cas9 variants using sgRNAs with matched 5' nucleotides. *Genome Biology* **18**, 218 (2017).
31. Chavez, A. et al. Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* **12**, 326–328 (2015).
32. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
33. Zhang, Y. et al. Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Sci. Rep.* **4**, 5405 (2014).
34. Kim, Y. B. et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* **35**, 371–376 (2017).
35. Komor, A. C. et al. Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci. Adv.* **3**, eaao4774 (2017).
36. Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by *in vitro* selection. *Nat. Methods* **8**, 765–770 (2011).
37. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
38. Kleinstiver, B. P. et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
39. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Badran, B. Hubbard, J. Levy, T. Huang, G. Church, A. Chavez, K. Esvelt, S. Vora and J. Scheiman for discussions. This work was supported by DARPA HR0011-17-2-0049, US NIH RM1 HG009490, R01 EB022376 and R35 GM118062 and the HHMI. J.H.H. was supported by NDSEG and NSF graduate fellowships. S.M.M. was supported by an NSF graduate fellowship. W.T. is an HHMI Fellow of the Jane Coffin Childs Memorial Fund. L.C. was supported by the Agency for Science, Technology, and Research, Singapore.

Author Contributions J.H.H. designed the research, performed PACE, characterized variants in bacteria, conducted human cell experiments, analysed data, performed off-target analysis and wrote the manuscript. S.M.M. performed human cell experiments, analysed data and wrote the manuscript. M.H.G. performed human cell experiments and data analysis. W.T. performed human cell experiments and cloning. L.C. and C.M.Z. assisted with cloning and PACE. N.S. optimized Cas9 PACE. X.G. assisted with off-target analysis. H.A.R. assisted with indel and base-editing analyses. Z.L. assisted with human cell experiments. D.R.L. designed and supervised the research and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.R.L. (drliu@fas.harvard.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

General methods and cloning. DNA sequences used in this work are listed in the Supplementary Information. PCR was performed using Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs) or Phusion U Green Multiplex PCR Master Mix (Thermo Fisher Scientific). PACE plasmids and phages were constructed by USER cloning or Gibson cloning (New England Biolabs). *Cas9* genes and plasmid backbones for PACE were obtained from previously reported plasmids^{19,20} that are available from Addgene. Plasmids encoding *dxCas9-VPR*³¹, *xCas9* nucleases, *xCas9-BE3*⁴, *xCas9-BE4*³⁵ and *xCas9-ABE*⁵ were constructed by replacing *SpCas9* using Gibson cloning. Plasmids for sgRNA expression were constructed using one-piece blunt-end ligation of a PCR product containing a variable 20-nucleotide sequence corresponding to the desired sgRNA targeted site. Primers and templates used in the synthesis of all sgRNA plasmids used in this work are listed in Supplementary Tables 8, 9, 11, 12, 15 and 16. PCR was performed using Q5 Hot Start High-Fidelity Polymerase (New England Biolabs) with phosphorylated primers and the plasmid pFYF1320 (sgRNA expression plasmid with a spacer for targeting enhanced GFP (eGFP)) as a template according to the manufacturer's instructions. PCR products were analysed by agarose gel electrophoresis, the band of the expected molecular weight was excised, and the DNA was extracted using a ZymoClean Gel DNA Recovery Kit (Zymo Research) and ligated using T4 DNA Ligase (New England Biolabs) according to the manufacturer's instructions. DNA vector amplification was carried out using Mach1 competent cells (Thermo Fisher Scientific). All mammalian ABE constructs, sgRNA plasmids and bacterial constructs were transformed and stored as glycerol stocks at -80°C in Mach1 T1^R Competent Cells (Thermo Fisher Scientific), which carry a *recA* mutation. Molecular biology grade Hyclone water (GE Healthcare Life Sciences) was used in all assays and PCR reactions. All vectors used in evolution experiments and mammalian cell assays were purified using ZymoPURE Plasmid Midiprep (Zymo Research), which includes endotoxin removal. Antibiotics were purchased from Gold Biotechnology.

Cell culture. HEK293T (ATCC CRL-3216) and U2OS (ATCC-HTB-96) cells were maintained in DMEM plus GlutaMax (Thermo Fisher Scientific) supplemented with 10% (v/v) fetal bovine serum (FBS) at 37°C with 5% CO_2 . Cell lines tested negative for mycoplasma contamination.

Transfections. HEK293T cells were seeded on 48-well poly-D-lysine-coated BioCoat plates (Corning) and transfected at approximately 85% confluency. For genomic DNA cutting or base editing, 750 ng of *Cas9* or BE3 and 250 ng of sgRNA expression plasmids were transfected using 1.5 μl of Lipofectamine 2000 (Thermo Fisher Scientific) per well according to the manufacturer's protocol. For GFP activation, 200 ng of *dCas9-VPR* plasmid, 50 ng of sgRNA expression plasmid, 60 ng of GFP reporter plasmid and 30 ng of near-infrared fluorescent protein (iRFP) expression plasmid were transfected using 1.5 μl Lipofectamine 2000 (Thermo Fisher Scientific) per well according to the manufacturer's protocol. Endogenous gene activation was performed similarly but with 200 ng of *dCas9-VPR* plasmid and 50 ng of sgRNA expression plasmid only.

GFP transcriptional activation assay. Transfected HEK293T cells were trypsinized and resuspended in DMEM plus GlutaMax (Thermo Fisher Scientific) supplemented with 10% (v/v) FBS. The cells were kept on ice and flow cytometry was performed using a LSRII Fortessa from BD Biosciences. Events were gated for iRFP-positive cells to analyse transfected cells. The percentage of GFP-positive cells and the intensity of GFP fluorescence from each cell was collected.

RNA expression quantification for endogenous transcriptional activation assay. RNA was extracted from HEK293T cells using the Quick-RNA Plus Kit (Zymo Research). cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad) and quantitative PCR (qPCR) was performed on a Bio-Rad CFX96 Real-Time PCR Detection System using Q5 Polymerase (New England Biolabs) and SYBR Green (Lonza). Primers for qPCR are listed in Supplementary Table 10.

PAM depletion assay. Electrocompetent NEB 10-beta cells (New England Biolabs) were electroporated with two plasmids. The first plasmid expresses *Cas9* (inducible with anhydrotetracycline, ATc), the sgRNA (inducible with arabinose) and a spectinomycin-resistance gene. The second plasmid contains the target protospacer and a kanamycin-resistance gene. After incubation with SOC outgrowth medium (New England Biolabs) for 1 h the bacteria were plated on agar plates containing both spectinomycin and kanamycin along with ATc and arabinose inducers. After

incubating overnight, the bacterial cells on the agar plates were scraped and the plasmids extracted using the ZymoPURE Plasmid Midiprep Kit (Zymo Research). The resulting post-selection DNA included all of the protospacer plasmids not cleaved by *Cas9*. The same region around the protospacer in the pre-selection library and the post-selection DNA was then amplified separately using NEBNext High-Fidelity PCR Polymerase (New England Biolabs) with the flanking HTS primer pairs that are listed in the Supplementary Table 7. The Illumina barcoding PCR reaction was assembled with NEBNext High-Fidelity PCR Polymerase. PCR products were purified by electrophoresis with a 2% agarose gel using a QIAquick Gel Extraction Kit, eluting with 15 μl of water. DNA concentration was quantified with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems) and sequenced on an Illumina MiSeq instrument according to the manufacturer's protocols.

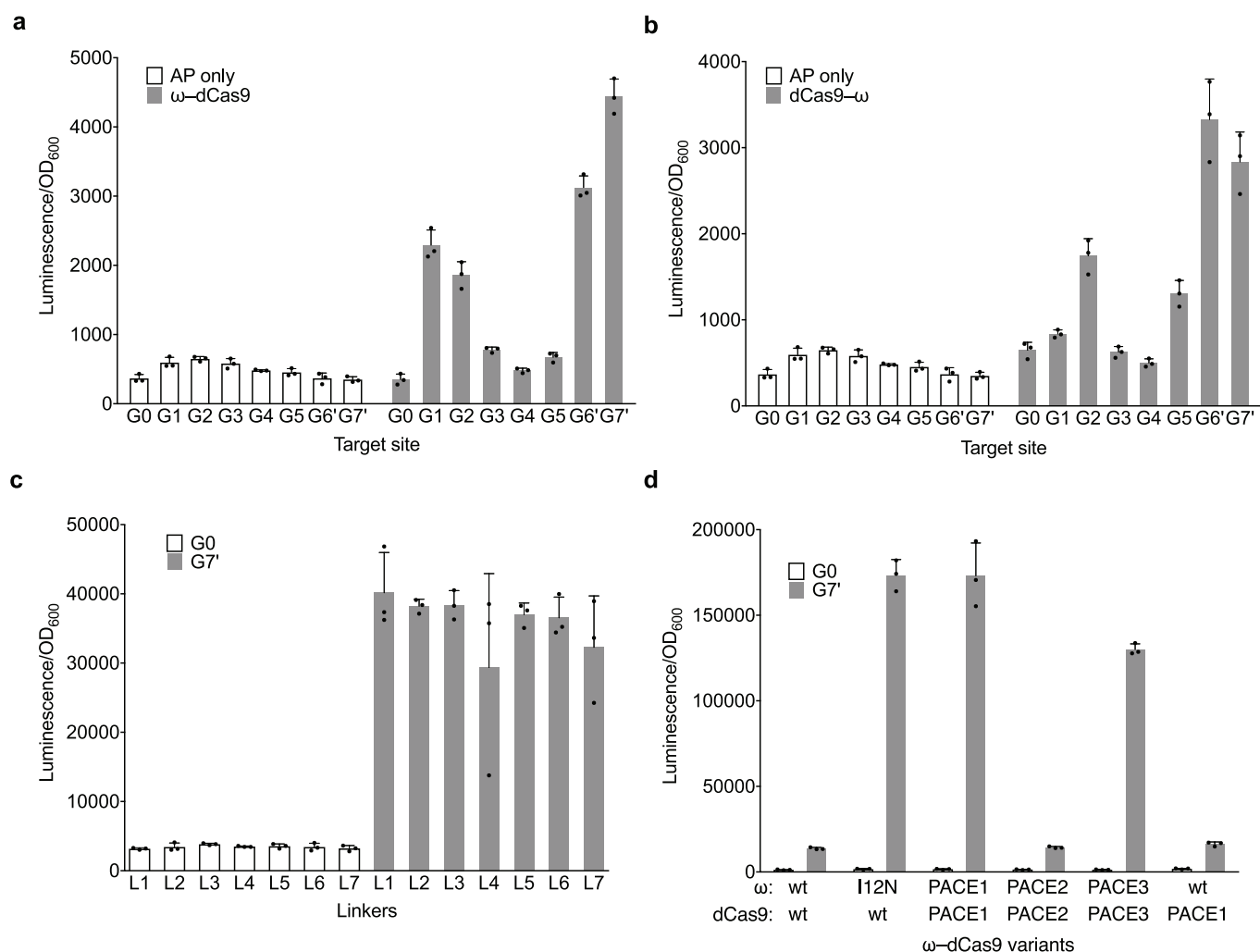
High-throughput DNA sequencing of genomic DNA samples. Transfected cells were harvested after 3 days (BE3 and BE4) or 5 days (DNA cutting and ABE). Medium was removed and cells were washed with $1\times$ PBS solution (Thermo Fisher Scientific). Genomic DNA was extracted by addition of 100 μl freshly prepared lysis buffer (10 mM Tris-HCl, pH 7.0, 0.05% SDS, 25 $\mu\text{g ml}^{-1}$ proteinase K (Thermo Fisher Scientific)) directly into each well of the tissue culture plate. The plate was incubated at 37°C for 1 h. The genomic DNA mixture was transferred to a 96-well PCR plate and incubated at 80°C for 15 min to denature enzymes. Genomic regions of interest were amplified by PCR with flanking HTS primer pairs that are listed in Supplementary Table 13. Each 25 μl PCR reaction was assembled with Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Fisher Scientific) according to the manufacturer's instructions using 1.0 μM forward and reverse primer and 1 μl of genomic DNA extract. PCR reactions were carried out under the following conditions: 95°C for 3 min, then 30 cycles of 98°C for 30 s, 60°C for 20 s and 72°C for 1 min, followed by a final 72°C extension for 5 min. PCR products were verified by comparison to DNA standards (1-kb Plus DNA Ladder) on a 2% agarose gel with ethidium bromide. Each 25- μl Illumina barcoding PCR reaction was assembled with Phusion DNA polymerase according to the manufacturer's instructions using 0.5 μM unique forward and reverse Illumina barcoding primer pair and 1 μl of unpurified genomic amplification PCR reaction mixture. The barcoding reactions were carried out as follows: 98°C for 2 min, then 8 cycles of 98°C for 12 s, 61°C for 25 s and 72°C for 30 s, followed by a final 72°C extension for 1.5 min. PCR products were purified by electrophoresis with a 2% agarose gel using a QIAquick Gel Extraction Kit, eluting with 15 μl of water. DNA concentration was determined with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems) and sequenced on an Illumina MiSeq instrument according to the manufacturer's protocols. Analysis was carried out using previously published Matlab code⁵ that is provided in Supplementary Notes 1 and 2.

Analysis of human disease-associated mutations in ClinVar database. Bioinformatic analysis of the ClinVar database was carried out in a manner similar to previously described analysis³⁴. The code is provided in Supplementary Note 3.

GUIDE-seq. HEK293T cells were transfected with 750 ng of the *Cas9* plasmid, 250 ng of the sgRNA plasmid and 20 pmol of GUIDE-seq dsODN. U2OS cells were transfected with 750 ng of the *Cas9* plasmid, 250 ng of the sgRNA plasmid and 100 pmol of GUIDE-seq dsODN. For both cell types, 20 μl of solution SE (Lonza) was used along with a Lonza Nucleofector 4-D. Program CM-137 was used for HEK293T cells and program DN-100 was used for U2OS cells. Genomic DNA was extracted using the Quick-DNA Miniprep Plus Kit (Zymo Research) following the manufacturer's protocol. The DNA was sheared to an average of 500 base pairs using a Covaris S220 focused ultrasonicator as previously described²⁹. End repair, dA-tailing, adaptor ligation, tag-specific PCR1 and tag-specific PCR2 were carried out using the primers and methods described previously²⁹. DNA concentration was quantified with the KAPA Library Quantification Kit-Illumina (KAPA Biosystems) and sequenced on an Illumina MiSeq instrument according to the manufacturer's protocols. Analysis was carried out using previously published Python code²⁹.

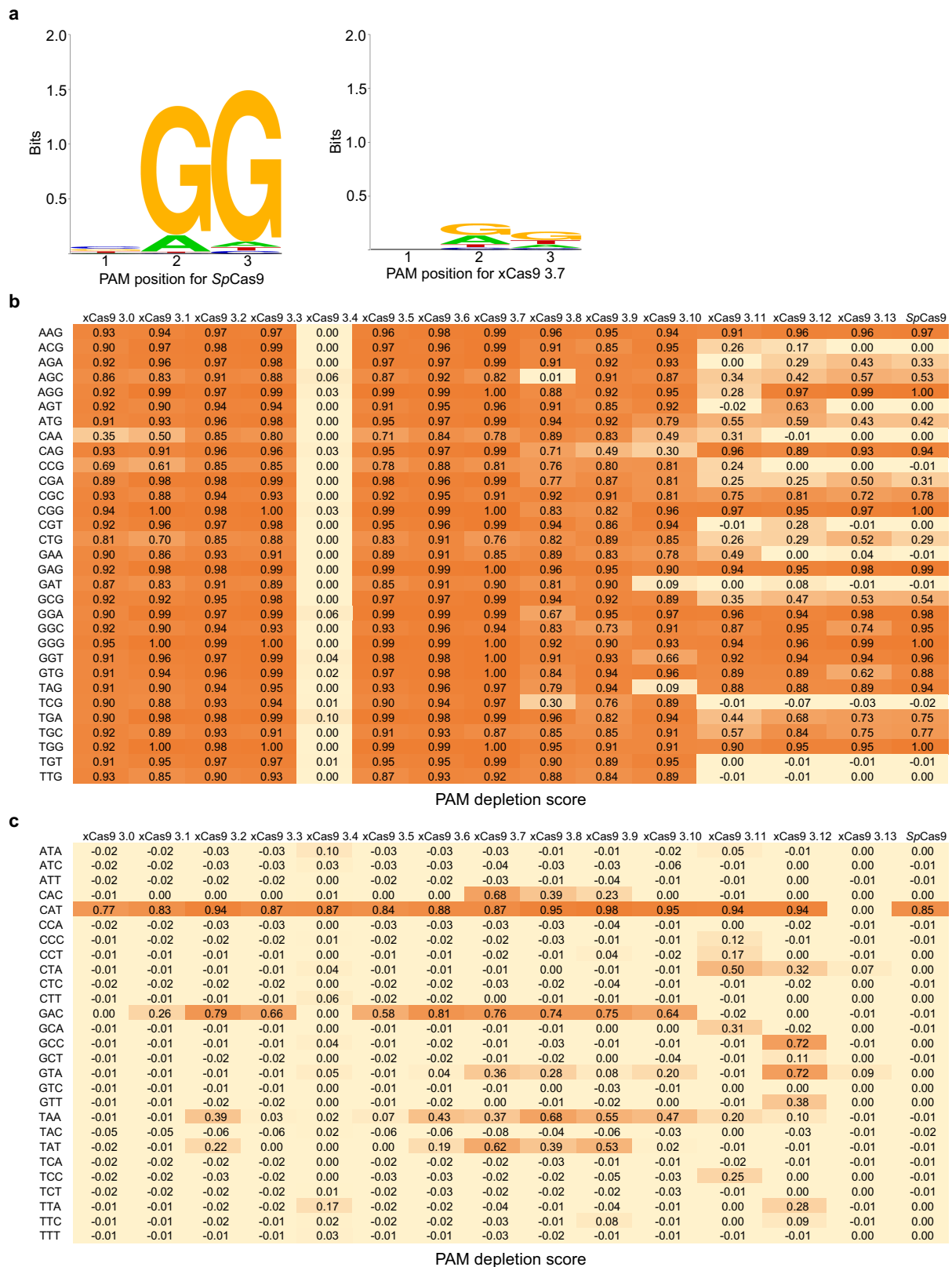
Data availability. High-throughput sequencing data have been deposited in the NCBI Sequence Read Archive database under accession code SRP130166. GUIDE-seq sequencing files listed in Supplementary Table 17. Plasmids encoding the *xCas9-3.7* and *-3.6* transcriptional activators, nucleases, BE3, BE4 and ABE have been deposited with Addgene.

40. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).



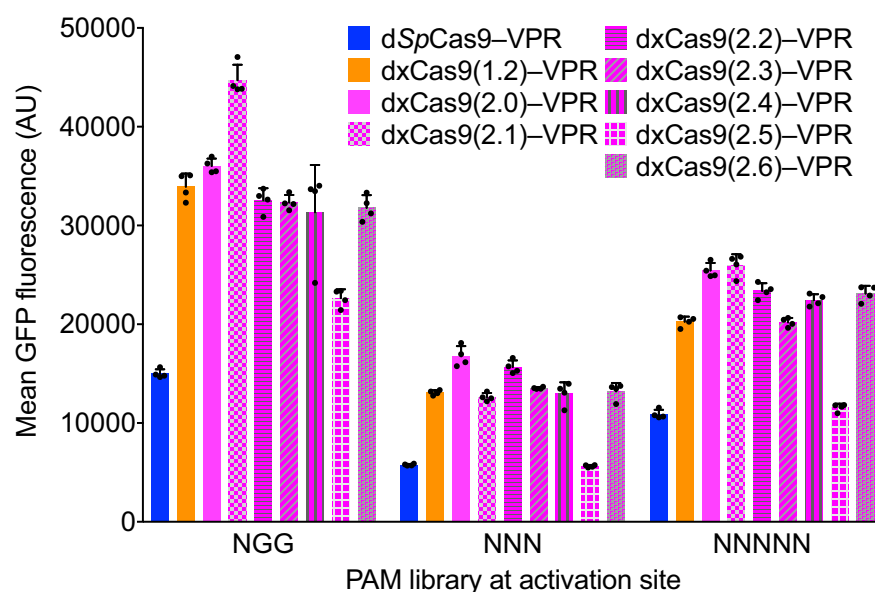
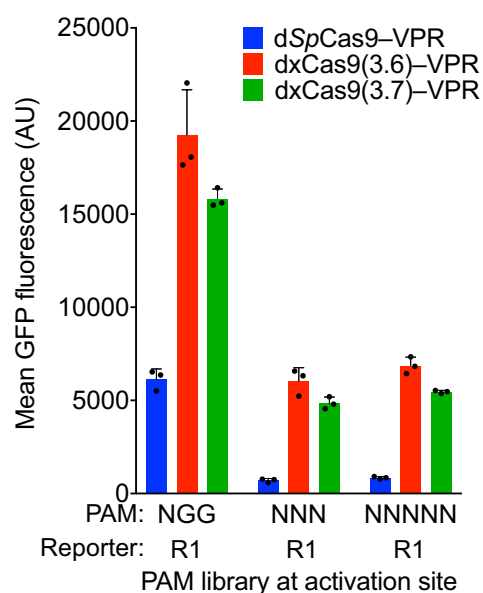
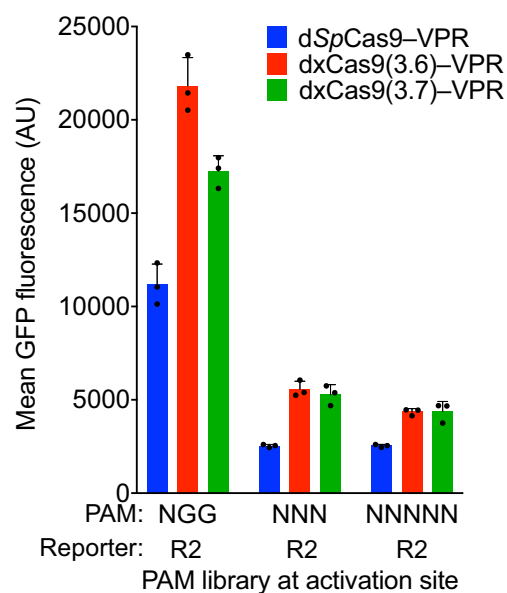
Extended Data Figure 1 | Optimization of Cas9 PACE. Luciferase expression in *E. coli* was used as a proxy of gene III expression levels during efforts to link Cas9 binding to gene expression for PACE. **a, b**, Seven guide RNAs targeting the luciferase reporter (G1–G7', see Supplementary Table 1), as well as a scrambled guide RNA negative control (G0) were tested without dCas9 (white bars) and with ω-dCas9 (**a**) or dCas9-ω (**b**) fusions (grey bars). **c**, Tests of seven different linkers

between ω and dCas9. See Supplementary Table 2 for linker sequences. **d**, Evolution of ω-dCas9 on an NGG PAM site in PACE yielded variants (PACE1, PACE2 and PACE3) that were tested in comparison with canonical wild-type (wt) ω-dCas9, ω tethered to PACE1 dCas9 and the I12N ω mutant tethered to canonical wild-type dCas9. **a–d**, Data are mean and s.d. of three biologically independent samples.



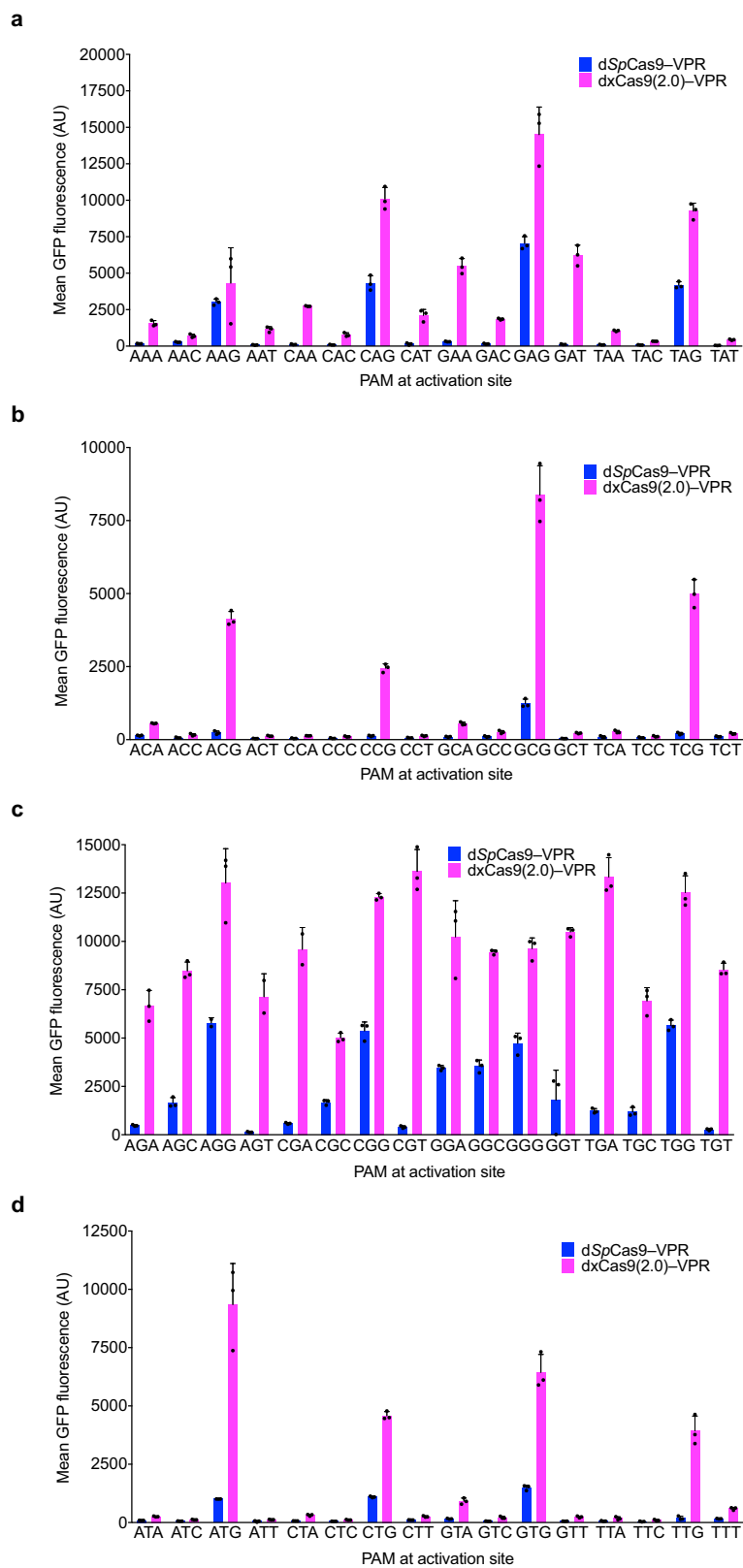
Extended Data Figure 2 | PAM profiling of xCas9 variants. **a**, In separate experiments, a plasmid library containing a protospacer with all possible NNN PAM sequences and a spectinomycin-resistance gene was electroporated into *E. coli* along with a plasmid expressing SpCas9 or the xCas9 variant shown. PAMs that are cleaved are depleted from the library when plated on medium containing spectinomycin. HTS of the library before and after selection enables quantification of the change in library

composition, resulting in a sequence logo⁴⁰ for the PAM preference of SpCas9 (left) and xCas9-3.7 (right). **b**, **c**, PAM depletion scores of Cas9 variants from spectinomycin selection in *E. coli*, calculated as described previously³⁶, with 1.0 representing complete cleavage of that PAM sequence. Scores for NGN, NNG, GAA, GAT and CAA are shown in **b** and the rest of the PAM sequences are shown in **c**.

a**b****c**

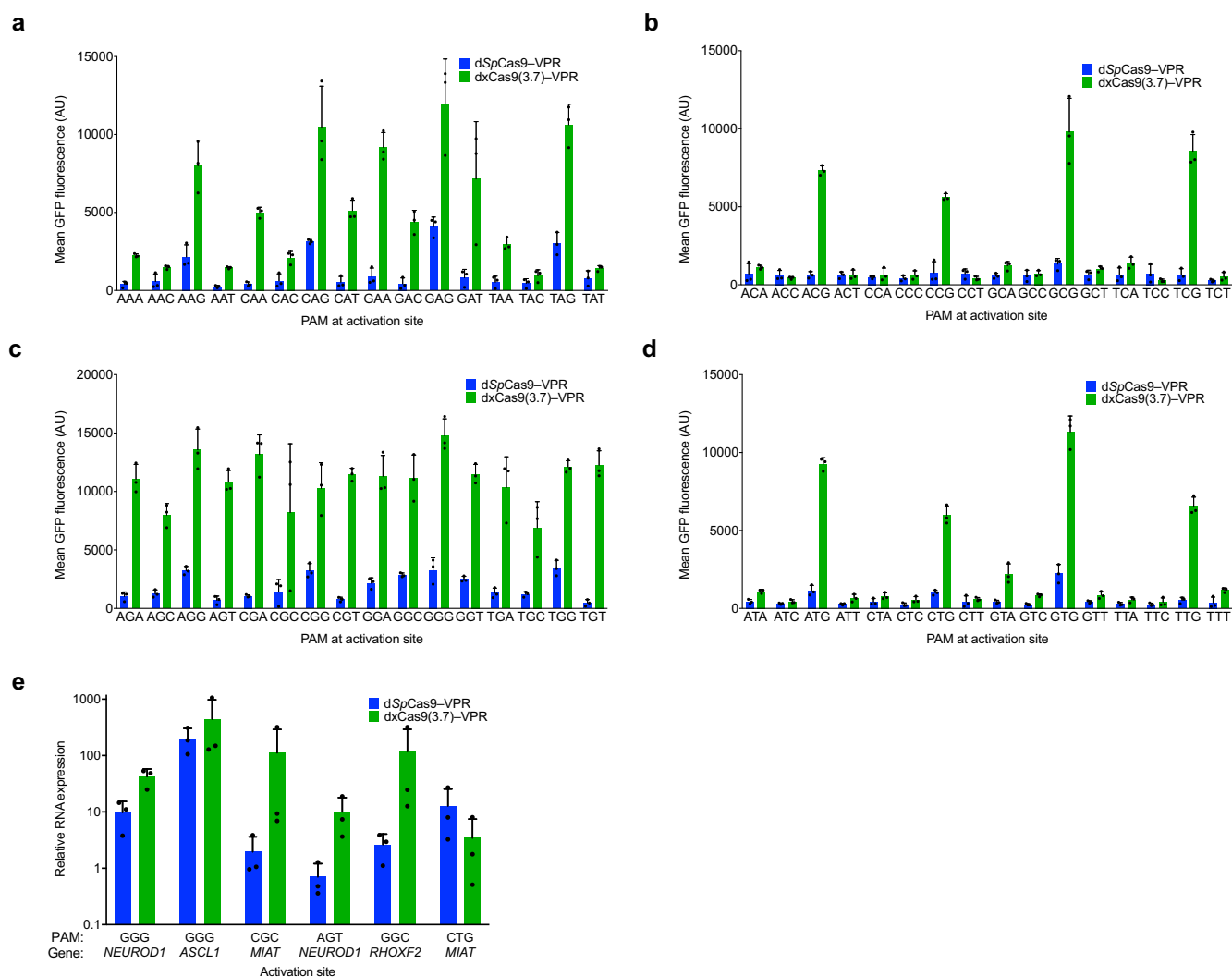
Extended Data Figure 3 | Transcriptional activation of reporter site PAM libraries with xCas9. Transcriptional activation by dSpCas9-VPR and dxCas9-VPR variants (transfected as plasmids) on GFP reporter plasmids containing different PAM sites in HEK293T cells. **a**, Earlier generations of xCas9 variants were tested on the R1 site with the NGG,

NNN or NNNNN PAM libraries. **b**, **c**, Transcriptional activators dxCas9(3.6)-VPR and dxCas9(3.7)-VPR were tested on two different protospacer reporters, R1 (**a**) and R2 (**b**), containing adjacent NGG, NNN, or NNNNN PAM libraries. **a**–**c**, Data are mean and s.d. of three biologically independent samples.



Extended Data Figure 4 | Transcriptional activation with xCas9-2.0.
a–d, The transcriptional activator dxCas9(2.0)-VPR was tested on the R1 protospacer (Extended Data Fig. 3 and Supplementary Table 8)

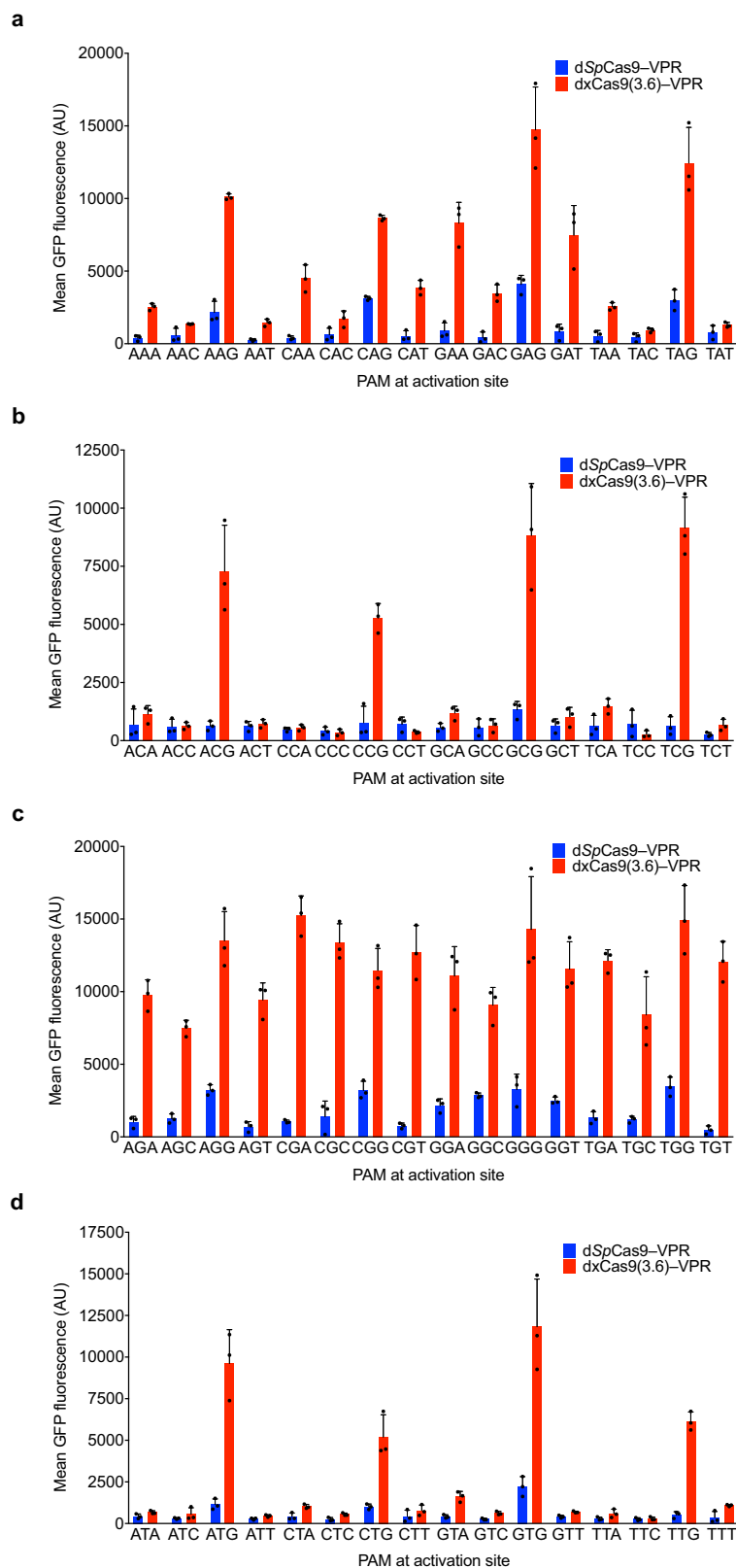
with each of the 64 possible NNN PAMs (NAN, NCN, NGN and NTN) in HEK293T cells. Data are mean and s.d. of three biologically independent samples.



Extended Data Figure 5 | Transcriptional activation with xCas9-3.7 on all 64 NNN PAM sites and endogenous gene activation in human cells.

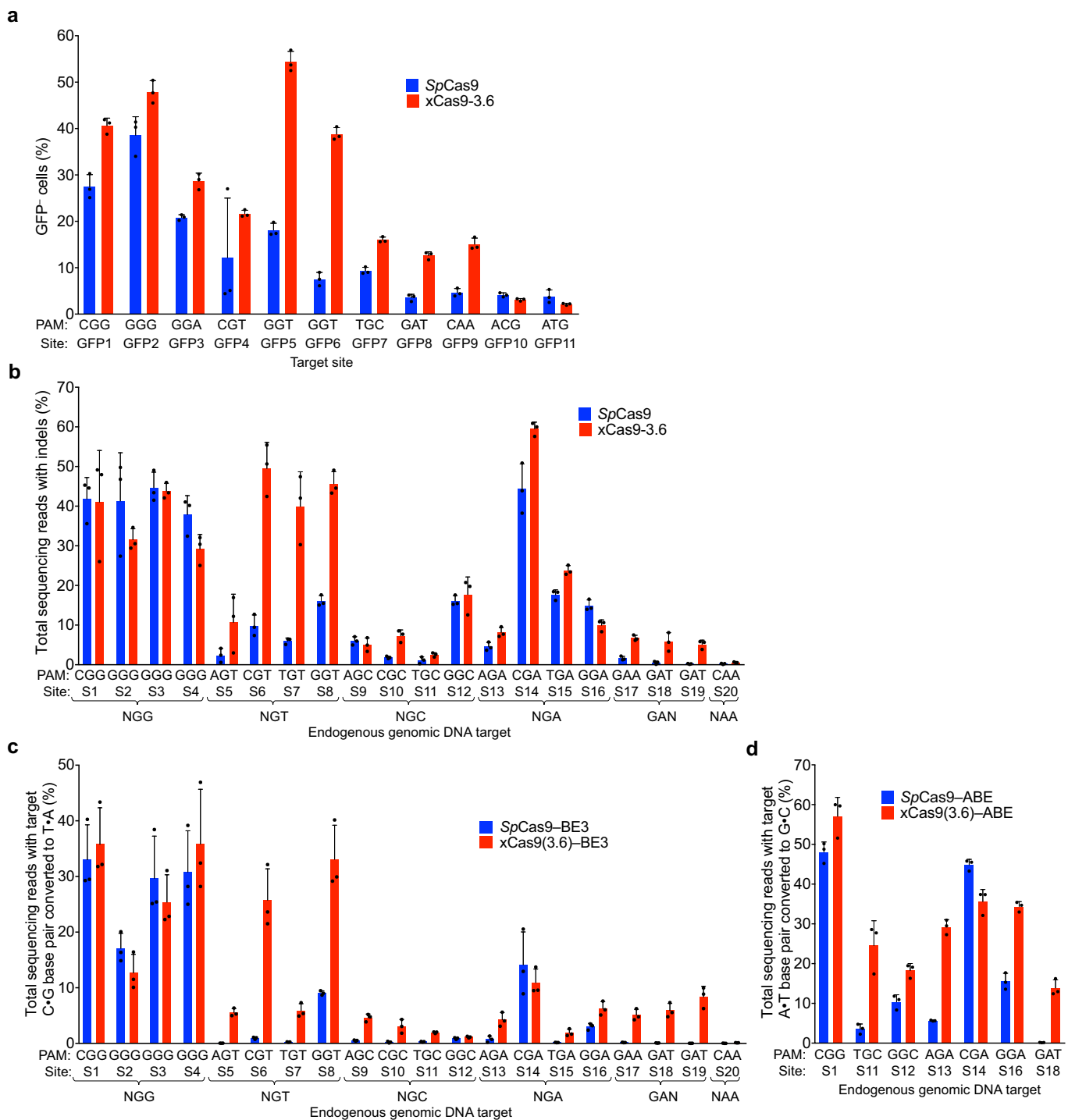
a–d, The transcriptional activator dxCas9(3.7)-VPR was tested on the R1 protospacer (Extended Data Fig. 3 and Supplementary Table 8) with each of the 64 possible NNN PAMs (NAN, NCN, NGN and NTN) in HEK293T cells. **e**, Endogenous gene activation was tested using both dSpCas9-VPR and dxCas9(3.7)-VPR to activate expression of the *NEUROD1*, *ASCL1*,

MIAT or *RHOXF2* at six total sites. RNA expression as measured by qPCR with reverse transcription (RT-qPCR) was compared to background expression levels for each gene (measured in the control with no sgRNA) and was normalized to *ACTB* expression. **a–e**, Data are mean and s.d. of three biologically independent samples. Target sites are listed in Supplementary Table 9.



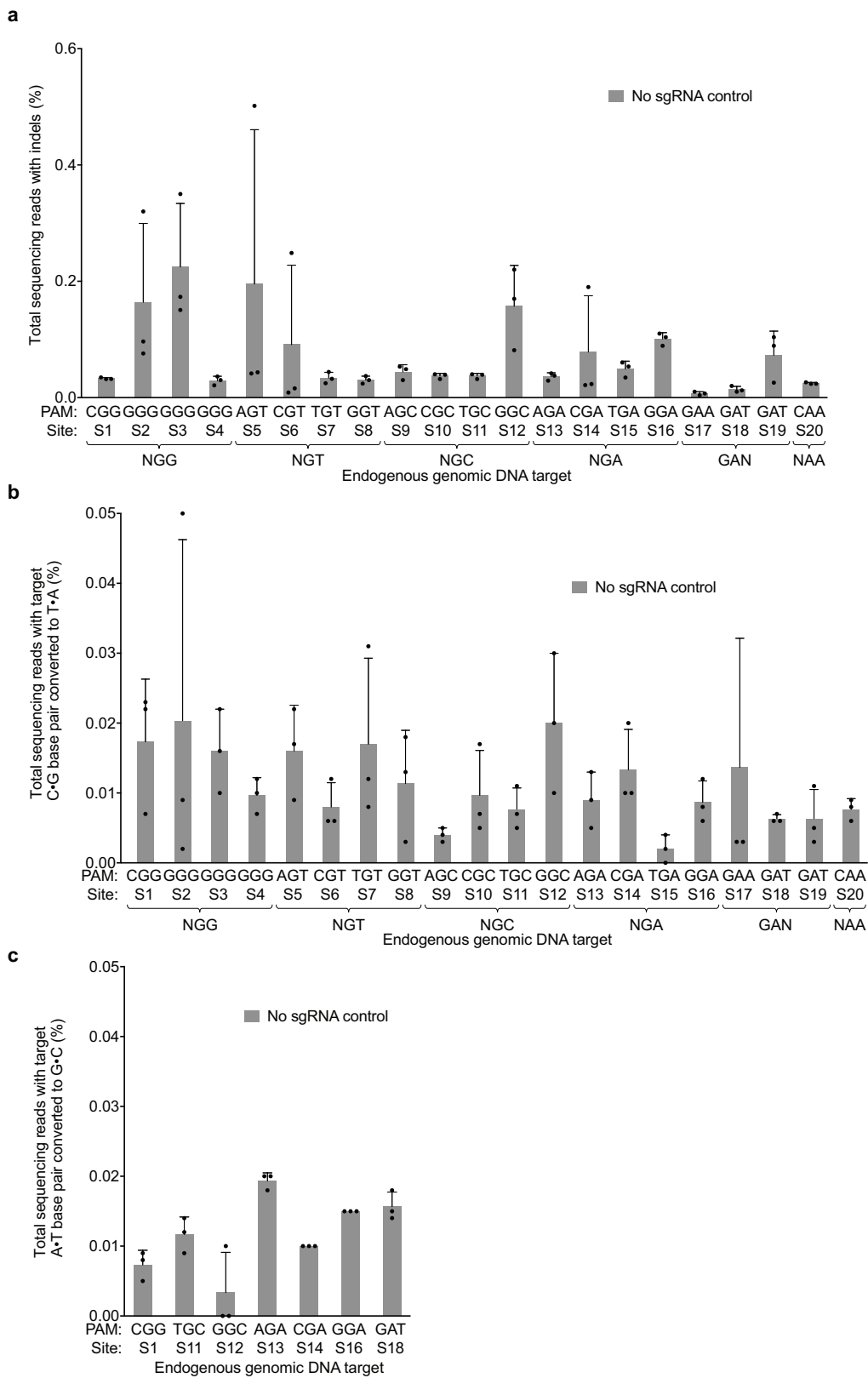
Extended Data Figure 6 | Transcriptional activation with xCas9-3.6 on all 64 NNN PAM sites. a–d, The transcriptional activator dxCas9(3.6)-VPR was tested on the R1 protospacer (Extended Data Fig. 3 and

Supplementary Table 8) with each of the 64 possible NNN PAMs (NAN, NCN, NGN and NTN) in HEK293T cells. Data are mean and s.d. of three biologically independent samples.



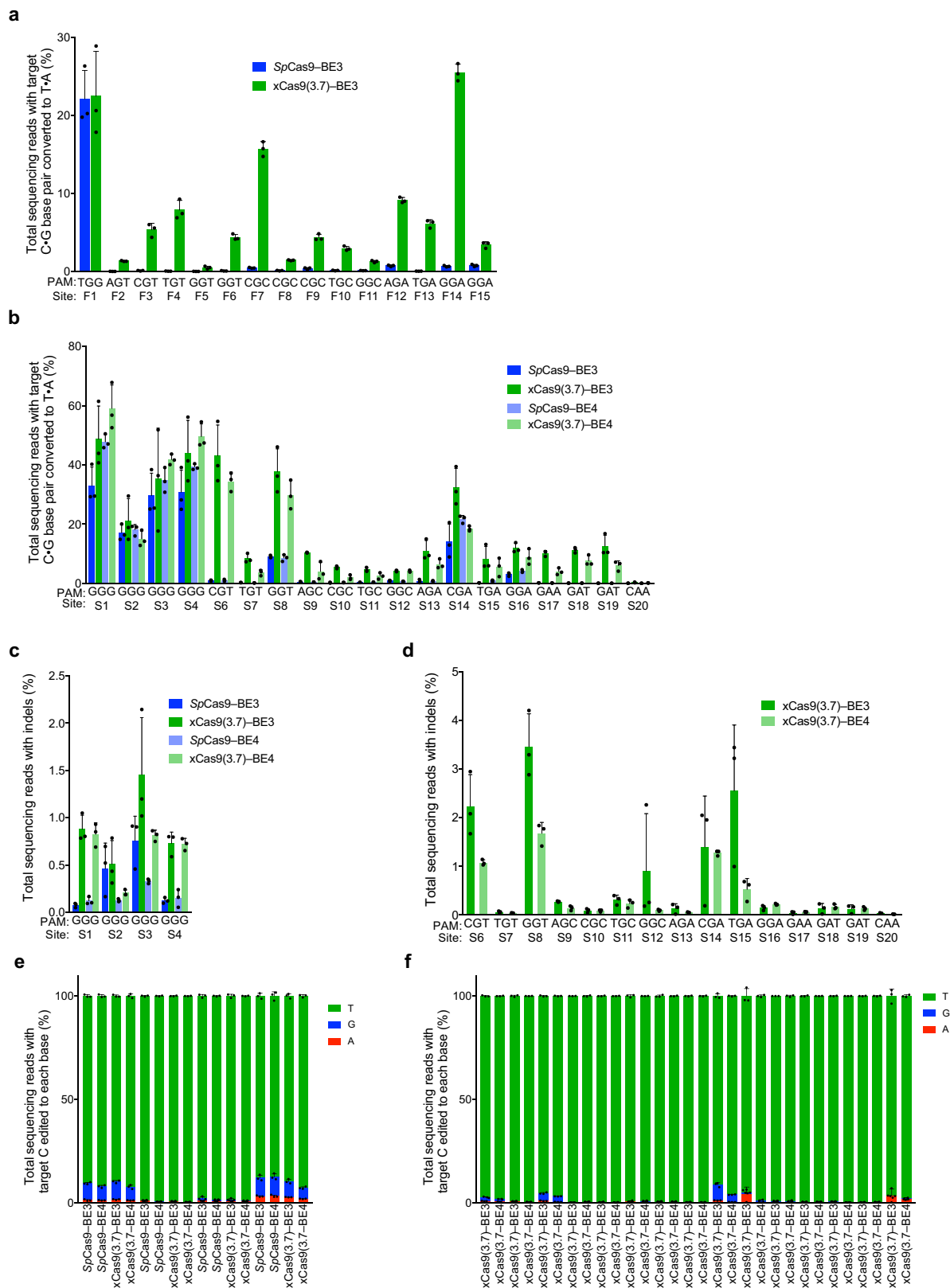
Extended Data Figure 7 | Genomic DNA cleavage and base editing by evolved xCas9-3.6. **a**, Genomic DNA cleavage by SpCas9 or xCas9-3.6 (transfected as plasmids), in HEK293-GFP cells containing a genomically integrated GFP gene. After 5 days, the cells were analysed for loss of GFP fluorescence by flow cytometry. Sequences for all target sites are listed in Supplementary Table 11. **b**, DNA cleavage of endogenous genomic DNA sites with a variety of NGG and non-NGG PAMs by SpCas9 and xCas9-3.6 in HEK293T cells. Indel rates were measured using HTS 5 days after plasmid transfection. Sequences for all target sites are listed in Supplementary Table 12. **c**, C•G-to-T•A base editing in HEK293T cells

by SpCas9-BE3 or xCas9(3.6)-BE3 was tested at 20 sites containing NG, GAA, GAT, or CAA PAM sites. The C•G-to-T•A conversion frequency, ascertained using HTS, at the most efficiently edited base 3 days after plasmid transfection is shown. **d**, Of the 20 sites in **c**, seven contained an A in the canonical window for ABE editing⁵ and were tested for A•T-to-G•C base editing by SpCas9-ABE and xCas9(3.6)-ABE. The A•T-to-G•C conversion frequency, ascertained using HTS, at the most efficiently edited base 5 days after plasmid transfection is shown. **a–d**, Data are mean and s.d. of three biologically independent samples. Complete HTS results across the protospacer are provided in Supplementary Table 14.



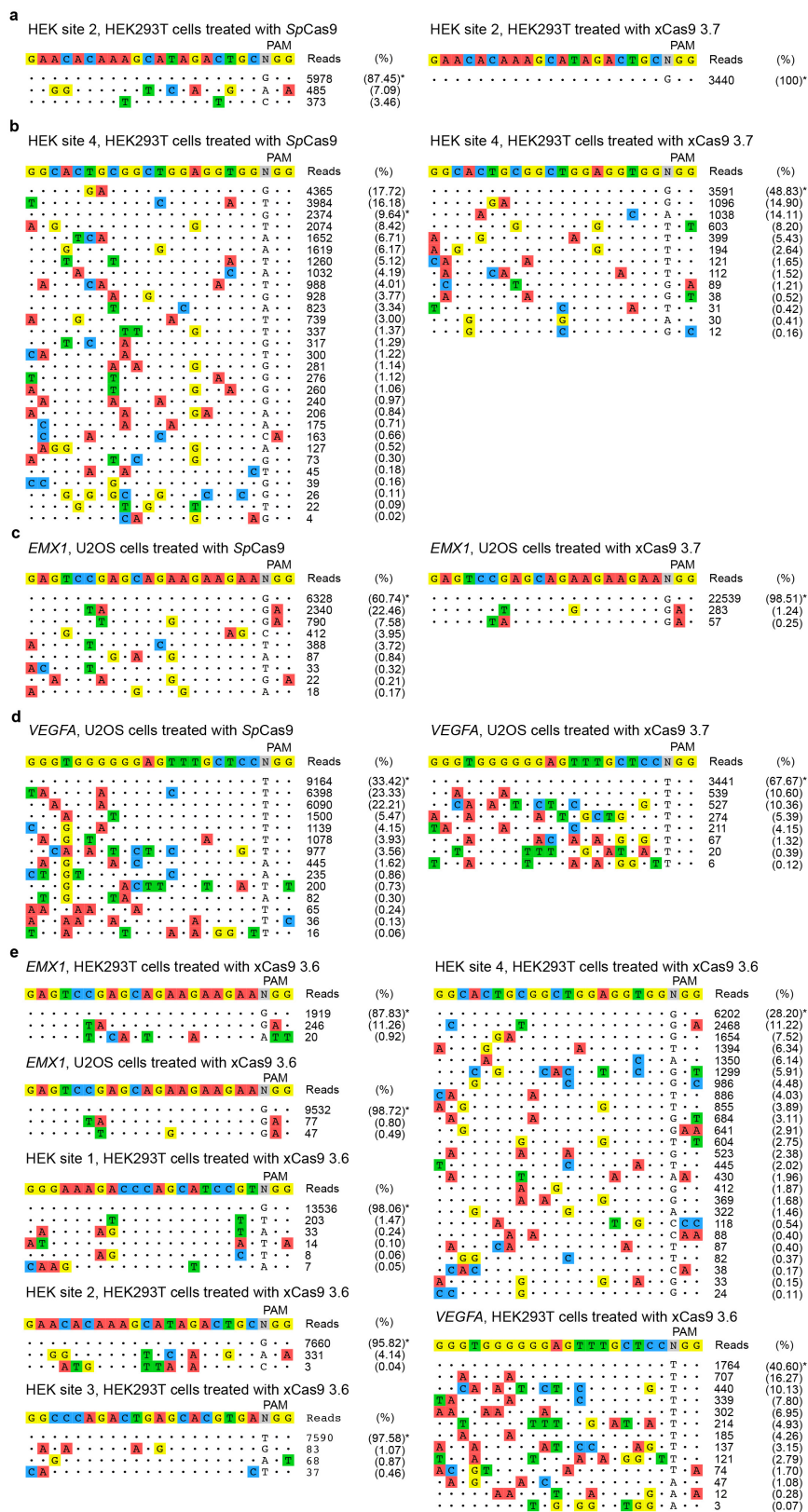
Extended Data Figure 8 | Negative controls lacking sgRNA for nuclease and base editing experiments. To verify genomic DNA cleavage and base editing results, the same sites were sequenced after treatment with *SpCas9* nuclease, *SpCas9*–BE3, or *SpCas9*–ABE but without any sgRNA. **a**, Indel rates at endogenous target sites 5 days after treatment of HEK293T cells

with SpCas9. **b**, Target C•G-to-T•A conversion 3 days after treatment of HEK293T cells with SpCas9-BE3. **c**, Target A•T-to-G•C base conversion 5 days after treatment of HEK293T cells with SpCas9-ABE. **a-c**, Data are mean and s.d. of three biologically independent samples. Complete HTS results across the protospacer are provided in Supplementary Table 14.



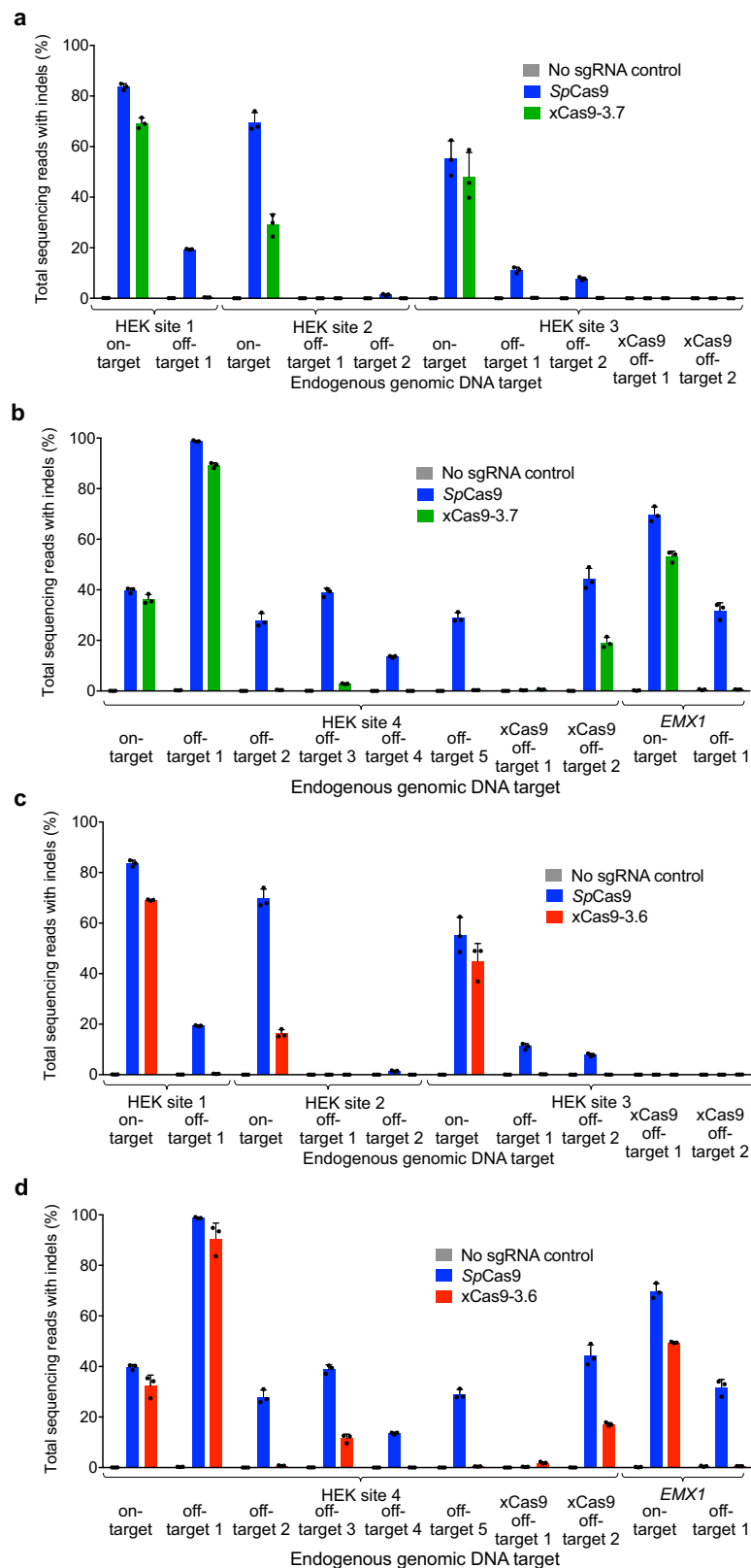
Extended Data Figure 9 | Cytidine base editing at 15 additional genomic sites and xCas9 base editing with the BE4 architecture. **a**, Base editing by SpCas9-BE3 and xCas9(3.7)-BE3 at 15 sites within the *FANCF* gene in HEK293T cells. The C•G-to-T•A conversion frequency at the most efficiently edited base 3 days after plasmid transfection is shown. **b**, Test of xCas9-3.7 in the BE4 architecture³⁵ on the same sites tested in Fig. 3. The C•G-to-T•A conversion frequency in HEK293T cells at the most efficiently edited base 3 days after plasmid transfection is shown. **c, d**, Indel frequency following treatment with BE3 or BE4 variants

targeting sites with NGG PAMs (**c**) and non-NGG PAMs (**d**). **e, f**, Product distribution among edited DNA sequence reads (reads in which the target C is mutated) following treatment with BE3 or BE4 variants targeting sites with NGG PAMs (**e**) and non-NGG PAMs (**f**). Because SpCas9 has minimal activity on non-NGG PAM sites, only xCas9(3.7)–BE3 and xCas9(3.7)–BE4 data are compared on non-NGG PAM sites. **a–f**, Data are mean and s.d. of three biologically independent samples. Target sites are listed in Supplementary Table 12.



Extended Data Figure 10 | Additional characterization of xCas9-3.7 and xCas9-3.6 by GUIDE-seq. a–d, In addition to the GUIDE-seq data shown in Fig. 4, two additional sites in HEK293T cells (a, b) and two sites in U2OS cells (c, d) were analysed after treatment with SpCas9 and xCas9-3.7. e, All six GUIDE-seq sites with an NGG PAM that were tested with

SpCas9 and xCas9-3.7 in HEK293T and U2OS cells were also tested with xCas9-3.6. On-target reads (indicated by an asterisk) and off-target reads for all sites are shown. Target sequences are listed in Supplementary Table 15.



Extended Data Figure 11 | Validation by high-throughput sequencing of GUIDE-seq results. a–d. The most frequent off-target sites identified using GUIDE-seq were verified by HTS of genomic DNA following treatment of HEK293T cells with SpCas9 or xCas9-3.7 (a, b), or following treatment with SpCas9 or xCas9-3.6 (c, d). New sites with non-NGG

PAMs that were identified as off-target sites of the xCas9 proteins were also analysed in a–d. a–d, Data are mean and s.d. of three biologically independent samples. Target sequences are listed in Supplementary Table 16.

Architecture of the human GATOR1 and GATOR1–Rag GTPases complexes

Kuang Shen^{1,2,3,4,*}, Rick K. Huang^{5,*}, Edward J. Brignole^{2,6}, Kendall J. Condon^{1,2,3,4}, Max L. Valenstein^{1,2,3,4}, Lynne Chantranupong^{1,2,3,4,†}, Aimaiti Bomaliyamu¹, Abigail Choe¹, Chuan Hong⁵, Zhiheng Yu⁵ & David M. Sabatini^{1,2,3,4}

Nutrients, such as amino acids and glucose, signal through the Rag GTPases to activate mTORC1. The GATOR1 protein complex—comprising DEPDC5, NPRL2 and NPRL3—regulates the Rag GTPases as a GTPase-activating protein (GAP) for RAGA; loss of GATOR1 desensitizes mTORC1 signalling to nutrient starvation. GATOR1 components have no sequence homology to other proteins, so the function of GATOR1 at the molecular level is currently unknown. Here we used cryo-electron microscopy to solve structures of GATOR1 and GATOR1–Rag GTPases complexes. GATOR1 adopts an extended architecture with a cavity in the middle; NPRL2 links DEPDC5 and NPRL3, and DEPDC5 contacts the Rag GTPase heterodimer. Biochemical analyses reveal that our GATOR1–Rag GTPases structure is inhibitory, and that at least two binding modes must exist between the Rag GTPases and GATOR1. Direct interaction of DEPDC5 with RAGA inhibits GATOR1-mediated stimulation of GTP hydrolysis by RAGA, whereas weaker interactions between the NPRL2–NPRL3 heterodimer and RAGA execute GAP activity. These data reveal the structure of a component of the nutrient-sensing mTORC1 pathway and a non-canonical interaction between a GAP and its substrate GTPase.

The mTORC1 pathway is a central regulator of cell growth^{1–5}. Nutrients signal to mTORC1 through the heterodimeric Rag GTPases (RAGA or RAGB bound to RAGC or RAGD)^{6–9}. When nutrients are abundant, RAGA binds GTP and RAGC binds GDP and this complex recruits mTORC1 to the lysosomal surface¹⁰, where RHEB stimulates the kinase activity of mTORC1^{11–16}. Upon nutrient starvation, the Rag GTPases adopt the opposite nucleotide loading state and cannot bind mTORC1, which becomes inhibited¹⁰.

The intrinsic GTP hydrolysis rates of the Rag GTPases are slow¹⁷, posing a problem for quickly altering the nucleotide state when nutrient levels change. Two GAP complexes, GATOR1^{18,19} and FLCN–FNIP2^{20,21}, have been discovered that stimulate GTP hydrolysis by RAGA or RAGB and RAGC or RAGD, respectively. Both GATOR1 and FLCN–FNIP2 are deregulated in human disease, with loss-of-function mutations in GATOR1 being a frequent cause of familial epilepsy^{22,23}.

GATOR1 has three stably interacting subunits: DEPDC5, NPRL2 and NPRL3. Despite its central role in mTORC1 signalling^{18,19,24}, there is currently an almost complete lack of structural information about this complex. Protein structure prediction software, such as I-TASSER²⁵ and Jpred²⁶, shows that all three subunits have low primary sequence similarity to other proteins and as a consequence have poorly defined domains. The only domains in GATOR1 with orthologous structures are two longin domains²⁷, one each at the N terminus of NPRL2 and NPRL3, and a DEP (Dishevelled, EGL-10 and pleckstrin) domain in DEPDC5. Here we used cryo-electron microscopy (cryo-EM) to solve the structure of GATOR1 on its own and in complex with the Rag GTPases.

Structural determination of GATOR1 complexes

To generate GATOR1 for structural studies we co-expressed NPRL2, NPRL3 and DEPDC5 in FreeStyle 293-F cells (Fig. 1a, b). To ensure a stable interaction between GATOR1 and the Rag GTPases, we purified a Rag GTPase heterodimer consisting of wild-type RAGA and

mutant RAGC(S75N) that eliminates its capacity to bind GTP but not GDP¹⁰. We loaded this heterodimer with an excess amount of guanosine 5′-[β,γ -imido]triphosphate (GppNHp; a non-hydrolysable GTP analogue) and GDP to lock its nucleotide-binding configuration to GppNHp•RAGA–RAGC(S75N)•GDP, which is the most favourable for interacting with GATOR1. Indeed, all five subunits co-eluted in the same fraction after gel filtration separation (Fig. 1a, b and Extended Data Fig. 1a, b). Consistent with previous studies^{17–19}, purified GATOR1 stimulated GTP hydrolysis by the RAGA–RAGC heterodimer by 14-fold, but had no effect on the complex containing mutant RAGA(Q66L) (Fig. 1c).

Well-defined particles of GATOR1 (290 kD) and the GATOR1–Rag GTPases complex (370 kD) were clearly visualized by cryo-EM (Extended Data Fig. 1c, d). Reference-free 2D classification revealed explicit structural details with views from different orientations (Extended Data Fig. 1e–g). High-resolution 3D refinements from a homogeneous subset of 3D classifications generated the final envelopes for GATOR1 (Fig. 1d) and the GATOR1–Rag GTPases complex (Fig. 1e) at 4.4 Å and 4.0 Å resolutions (gold-standard criteria, Fig. 1f).

Despite the lack of homologous structures for use as references, the electron microscopy density maps enabled direct tracing of backbones and registering of bulky residues, and thus the building of a tentative structural model for GATOR1 *de novo*. We resolved roughly 75% of GATOR1, except for two flexible regions in DEPDC5 that lack corresponding electron microscopy density (Fig. 2a, b). For the core region of DEPDC5, we reached near-atomic resolution where secondary structures and side chains were unambiguously resolved (Extended Data Fig. 1h–j). Within the GATOR1–Rag GTPases complex, GATOR1 adopts a similar conformation to that of free GATOR1. Because the Rag GTPase heterodimer shares sequence similarity with its yeast homologue, Gtr1p–Gtr2p^{28,29}, we were able to fit a homologous model into the extra electron microscopy density (Fig. 2c, d).

¹Whitehead Institute for Biomedical Research and Massachusetts Institute of Technology, Department of Biology, 455 Main Street, Cambridge, Massachusetts 02142, USA. ²Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ³Koch Institute for Integrative Cancer Research, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. ⁴Broad Institute of Harvard and Massachusetts Institute of Technology, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ⁵Howard Hughes Medical Institute, Janelia Research Campus, 19700 Helix Drive, Ashburn, Virginia 20147, USA. ⁶Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [†]Present address: Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA.

*These authors contributed equally to this work.

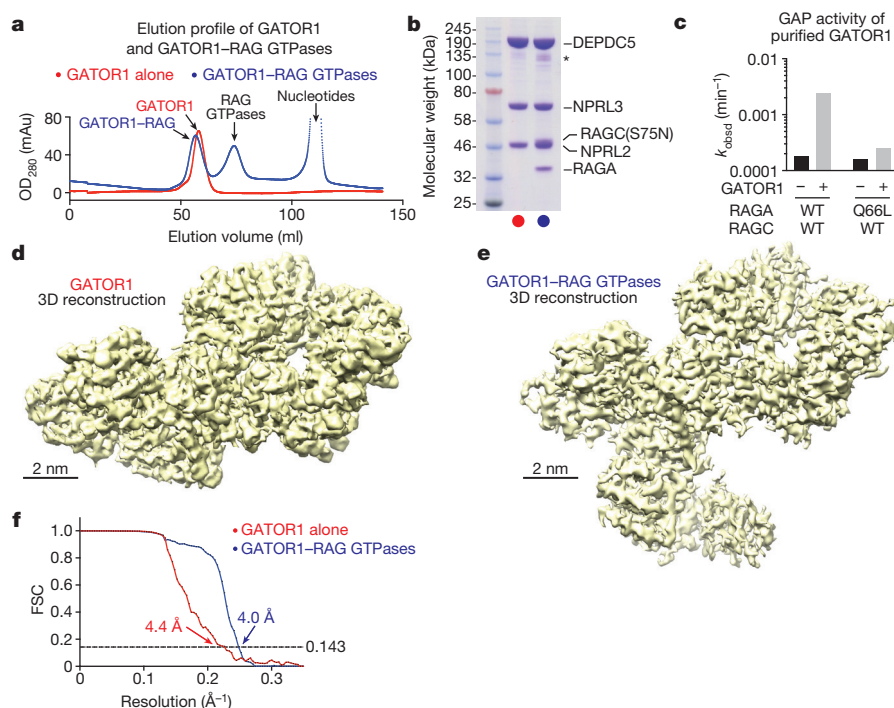


Figure 1 | Structural determination of GATOR1 and the GATOR1-Rag GTPases complex.

a, Gel filtration profiles for GATOR1 (red line) and GATOR1-Rag GTPases (blue line). mAU, milli-absorbance unit. **b**, Coomassie blue stained SDS-PAGE analysis of purified GATOR1 (red) and GATOR1-Rag GTPases (blue). Asterisk denotes a non-specific contamination that co-purifies with GATOR1. **c**, *In vitro* GAP activity of purified GATOR1. WT, wild type; k_{obsd} , observed rate constant. **d**, **e**, Envelopes of GATOR1 (**d**) and GATOR1-Rag GTPases (**e**) from the 3D reconstructions with density shown at 0.05 threshold level (UCSF Chimera). Scale bars, 2 nm. **f**, Gold-standard Fourier shell correlation (FSC) for GATOR1 (red line) and the GATOR1-Rag GTPases (blue line). Data in **a–c** are representative of two independent experiments. See Supplementary Table 1 for cryo-EM data collection and refinement.

Architecture of GATOR1 and GATOR1-Rag GTPases

The structural model reveals that the GATOR1 subunits contain several previously unidentified domains. DEPDC5 has five domains, which we named—in order from the N terminus to the C terminus—the N-terminal domain (NTD), structural axis for binding arrangement (SABA) domain, steric hindrance for enhancement of nucleotidase activity (SHEN) domain, DEP domain and C-terminal domain (CTD) (Fig. 2e and Extended Data Fig. 2a, b). Although the DEP domain is well-defined, to our knowledge the other four domains are here resolved and visualized for the first time.

The NTD localizes to the lateral side of DEPDC5 (Extended Data Fig. 2b). It has two lobes, both of which consist of a β -sheet with an adjacent α -helix (Extended Data Fig. 2c, d). A VAST search³⁰ for homologous structure models shows that lobe B shares structural similarity with the NTD of PEX1 AAA-ATPase (Extended Data Fig. 2e), which may serve as an adaptor for ubiquitin or UBX domains³¹.

The SABA domain (residues 168–427, previously annotated as DUF3608, domain of unknown function 3608) immediately follows the NTD of DEPDC5 (Fig. 3a). It has a globular shape and shares topological similarities with the NADP domain of flavodoxin reductase (NDFR)³² and the CD11a I-domain³³ (CD11I, Extended Data Fig. 2f–h), both of which contain ligand-binding motifs, NDFR for flavodoxin and CD11I for manganese(II). The SABA domain consists of six β -strands (β S1– β S4, β S6, β S9) that form a platform surrounded by four α -helices (α S1– α S4), two on each side (Fig. 3a). It is conserved at the sequence level in Im1p^{24,34}, the yeast homologue of DEPDC5, and organizes the assembly of GATOR1 by mediating interactions with the NPRL2–NPRL3 heterodimer (see below).

The SHEN domain (residues 720–1,010) connects to the SABA domain through a loop. Four β -strands construct its base, and two α -helices cover one side of the sheet (Fig. 3a). The SHEN domain uses two flexible regions (linker S and loop S) to form interdomain

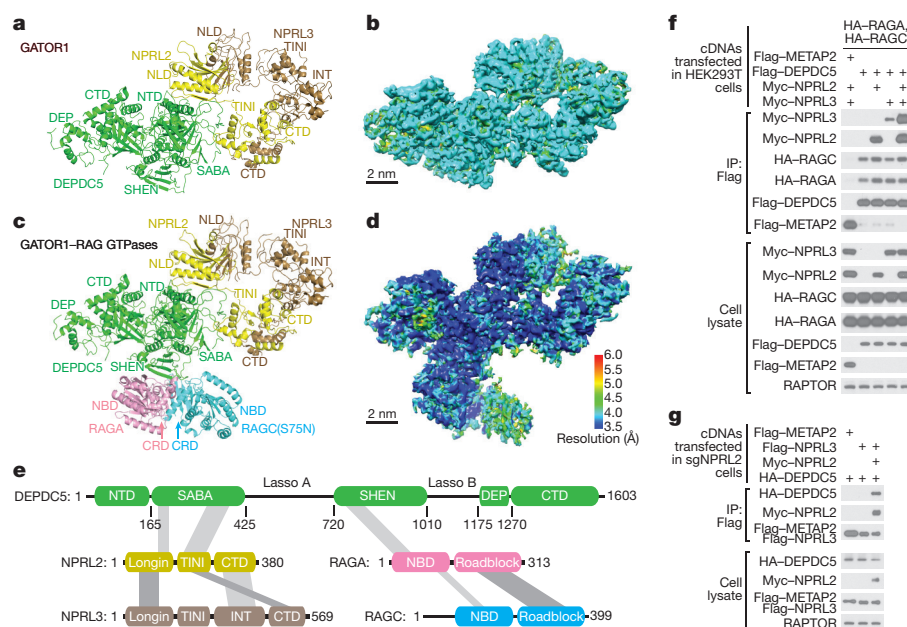


Figure 2 | Architectures of GATOR1 and the GATOR1-Rag GTPases complex.

a, **c**, Atomic models and domain assignment for GATOR1 (**a**) and the GATOR1-Rag GTPases complex (**c**). **b**, **d**, Local resolution of GATOR1 (**b**) and the GATOR1-Rag GTPases complex (**d**). **e**, Domain organization and interaction map for the GATOR1-Rag GTPases complex. Grey bars indicate domain–domain interactions. **f**, **g**, Co-immunoprecipitation assay to validate interactions amongst subunits of the GATOR1-Rag GTPases complex in wild-type HEK293T (**f**) and sgNPRL2 cells (**g**). Data in **f** and **g** are representative of two independent experiments. For gel source data, see Supplementary Figure 1. See Supplementary Table 1 for model building and validation.

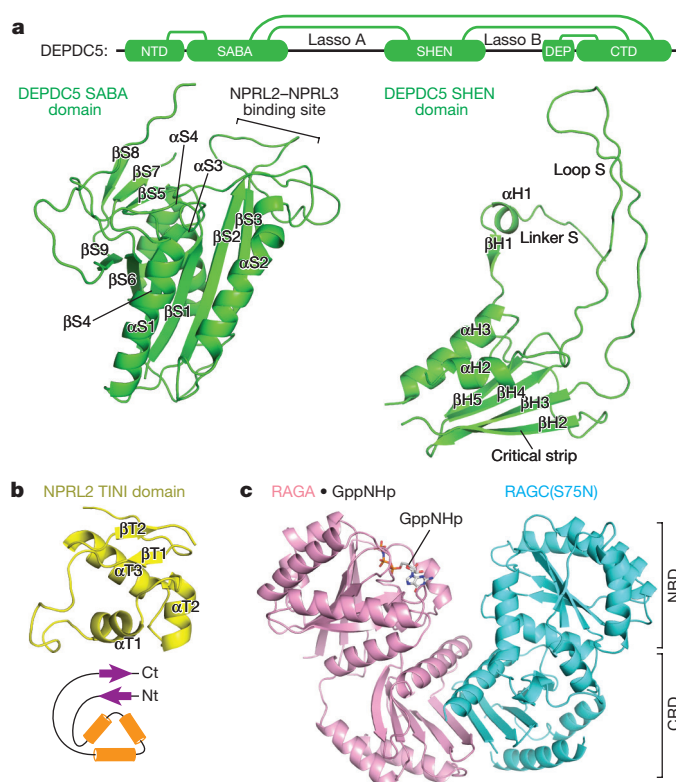


Figure 3 | Domain structures within the GATOR1–Rag GTPases complex. **a**, Structures for the SABA and SHEN domains of DEPDC5. Interdomain contacts are illustrated with green lines. **b**, Structure and topological diagram for the TINI domain of NPRL2. Nt, N terminus; Ct, C terminus. **c**, Structure of the RAGA–RAGC(S75N) heterodimer. NBD, nucleotide-binding domain. CRD, C-terminal roadblock domain.

contacts. Linker S contains a β -strand (β H1) and an α -helix (α H1). Notably, β H1 forms a continuous sheet with the β -strands in the NTD, inserting itself at the interface between the NTD and the SABA domain (Extended Data Fig. 3c). Loop S resides between α H2 and β H3 and directly contacts the SABA domain near where this domain binds with NPRL2–NPRL3, which could potentially mediate interdomain communication (Extended Data Fig. 3d, e). A β -strand (β H2), which we named the ‘critical strip’, contacts the nucleotide-binding domain of RAGA (Extended Data Fig. 3f, g). This interaction has a unique function and is indispensable for normal cellular response to amino acids, and thus differentiates the GATOR1–Rag GTPases from other GAP–GTPase pairs (see below).

The CTD (residues 1,291–1,603) of DEPDC5 contains two structurally similar lobes and has a pseudo-2-fold rotational symmetry (Extended Data Fig. 4a–c). Each half consists of a five-stranded β -sheet, with an α -helix covering one side. The CTD is located in the core of DEPDC5 and contacts all the other domains of DEPDC5 except the NTD, making it the central organizer of this multi-domain protein (Fig. 3a).

NPRL2 and NPRL3 have similar domain organizations (Extended Data Fig. 5a, f). Both contain an N-terminal longin domain (NLD, Extended Data Fig. 5b, g) that heterodimerizes (Extended Data Fig. 5k). After the NLD, a small domain bridges the longin domain to the C-terminal domains (Fig. 3b and Extended Data Fig. 5a). For NPRL2, this domain also mediates partial interactions with the SABA domain of DEPDC5 and we therefore renamed it the TINI domain (tiny intermediary of NPRL2 that interacts (with DEPDC5)). Besides the longin domain interactions, the C-terminal domains of NPRL2 and NPRL3 form a vast contact surface between each other that further reinforces their interaction (Extended Data Fig. 5k–m).

The Rag GTPase heterodimer shares a similar architecture with Gtr1p–Gtr2p (Fig. 3c). The N-terminal regions of RAGA and RAGC

contain the guanine-nucleotide binding domains (NBDs, Extended Data Fig. 6a). Within the nucleotide-binding pocket of RAGA we can clearly observe extra electron microscopy density corresponding to GppNHp (Extended Data Fig. 6b). The nucleotide-binding pocket of RAGC lacks sufficient resolution to identify the ligand that is bound, which is thought to be GDP. RAGA and RAGC heterodimerize via their C-terminal roadblock domains (CRD, Fig. 3c and Extended Data Fig. 6c), as has also been observed in other mTORC1 pathway components, such as the p14–MP1 heterodimer³⁵. Globally, the nucleotide-binding domains of RAGA and RAGC(S75N) are rotated substantially farther away from one another than seen in the open state of Gtr1p–Gtr2p (Extended Data Fig. 6d)^{28,29}, suggesting that regulation of this GTPase heterodimer might have diverged during evolution.

The structural model also revealed the interactions between the subunits. DEPDC5 directly contacts RAGA and NPRL2, NPRL3 is bound to NPRL2 and RAGC to RAGA. Co-immunoprecipitation experiments validated these conclusions: in the absence of other GATOR1 subunits, DEPDC5 can interact with NPRL2 and the Rag GTPases, and DEPDC5 co-immunoprecipitated NPRL3 to a much greater extent when NPRL2 was also co-expressed (Fig. 2f, g).

To identify the subunits of GATOR1 needed for it to associate with its known partners, we determined the capacity of GATOR1 subunits to co-immunoprecipitate endogenous GATOR2¹⁹, KICSTOR³⁶, and SAMTOR³⁷. Overexpressing DEPDC5 alone is sufficient to bind to KICSTOR and SAMTOR (Extended Data Fig. 7a), and NPRL3 is necessary and sufficient for the interaction with GATOR2 (Extended Data Fig. 7a, b). Because SAMTOR is a sensor for S-adenosylmethionine, and GATOR2 binds the leucine and arginine sensors, these results suggest that the nutrient availability is transmitted to GATOR1 through various interfaces (Extended Data Fig. 7c).

An intact GATOR1 is required for its GAP function

DEPDC5 interacts with NPRL2 through the SABA domain (Fig. 4a). Among the large number of residues at the contact surface (Extended Data Fig. 8a–d), we observed three loops on the tip of the SABA domain that directly contact NPRL2, which we defined as loops A (β S1– α S1, red in Fig. 4a), B (β S4– β S5, orange in Fig. 4a), and C (β S9–C terminal, blue in Fig. 4a). Specifically, loop A contacts a unique hairpin motif (Extended Data Fig. 5c) attached to the longin domain of NPRL2, whereas loop B and loop C contact the TINI domain of NPRL2 (Extended Data Fig. 8e–g). To investigate the roles of these contacts in mediating the DEPDC5–NPRL2 interaction, we generated DEPDC5 mutants in which these loops were mutated to flexible Gly-Ser(GS) linkers of the same length. Mutants replacing any one of the three loops had only a minor defect in binding NPRL2, as they still co-immunoprecipitated NPRL2 and NPRL3 in cells (Fig. 4b). However, we observed a strong synergistic effect when we replaced both loop A and loop B with GS linkers: the compound mutant in which both A and B loops were replaced (‘mutant AB’) failed to interact with any NPRL2 and NPRL3 (lane AB in Fig. 4b). These results suggest that loop A and B form redundant interactions with NPRL2 and are essential for forming an intact GATOR1 complex.

We next investigated whether an intact GATOR1 is necessary for the appropriate regulation of mTORC1 signalling by nutrients. In HEK293T cells lacking DEPDC5, mTORC1 signalling, as detected by the phosphorylation of its substrate S6K1, is higher than that in control cells and is largely resistant to amino acid starvation (Fig. 4c). Expression of wild-type DEPDC5 in these cells restores normal levels of mTORC1 signalling as well as its sensitivity to amino acids (Fig. 4c). By comparison, mutant AB fails to re-sensitize the DEPDC5-null cells to amino acid starvation (Fig. 4c). This result suggests that an intact GATOR1 is necessary for suppressing mTORC1 activity under nutrient-deficient conditions.

DEPDC5–Rag GTPases interaction is inhibitory

The SHEN domain of DEPDC5 directly contacts the NBD of RAGA (Fig. 5a). In our model, we resolve three pairs of hydrogen bonds

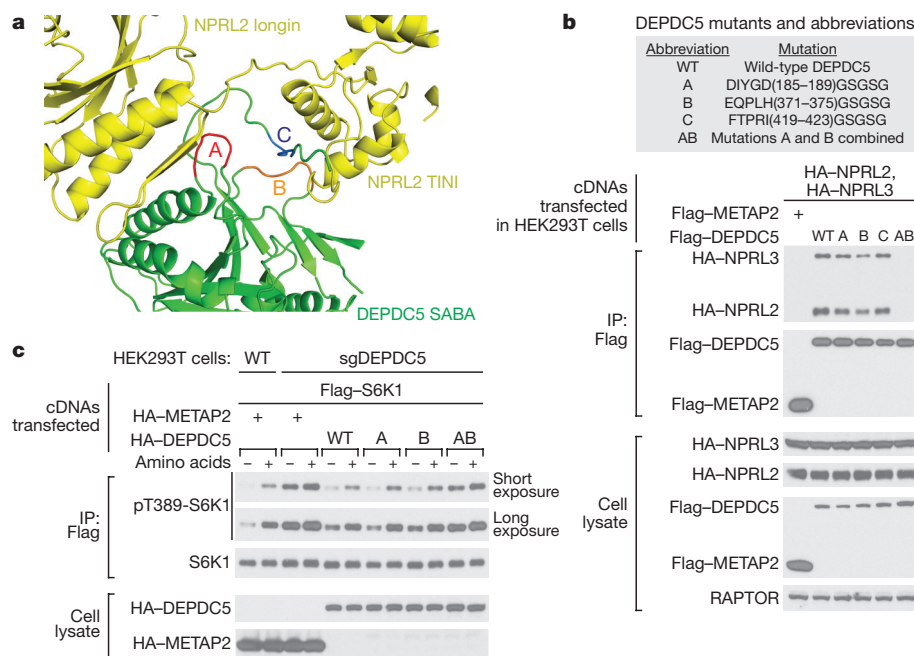


Figure 4 | An intact GATOR1 is necessary for mTORC1 inhibition upon amino acid starvation. **a**, Three loops in the SABA domain of DEPDC5 mediate the DEPDC5–NPRL2 interaction. Loops A, B and C are coloured in red, orange and blue, respectively. **b**, Compound mutant of loop A and loop B in DEPDC5 disrupts GATOR1 assembly. **c**, Expression of a DEPDC5 mutant that prevents GATOR1 assembly does not restore normal mTORC1 signalling in cells lacking DEPDC5. Data in **b** and **c** are representative of two independent experiments. For gel source data, see Supplementary Figure 1.

(Fig. 5b). Two of them are formed between RAGA and the backbone of the critical strip of DEPDC5, suggesting that the β -strand conformation of this segment of DEPDC5 may be crucial for mediating the interaction. We tested this possibility by investigating how variants of DEPDC5 with point mutations in the critical strip (residues 770–778) interact with the Rag GTPases in a co-immunoprecipitation assay. The DEPDC5(Y775A) mutant severely disrupted the interaction of DEPDC5 with the Rag GTPases (Fig. 5c and Extended Data Fig. 9a, b). Considering that the side chain of Tyr775 faces away from RAGA and that its backbone does not contact RAGA, we suspected that muta-

tion of this residue disrupts the conformation of the entire β -strand. Indeed, a much more severe mutation that we call ‘mutant P’—in which Tyr775–Pro779 of DEPDC5 was mutagenized to GS linkers (YDLLP(775–779)GSGSG)—did not further reduce the DEPDC5–Rag GTPases interaction compared with DEPDC5(Y775A) (Fig. 5c).

During GTP hydrolysis, canonical GAPs insert either an arginine finger or an asparagine thumb into the nucleotide-binding pocket of the target GTPase^{38,39}. We did not observe any extra electron microscopy density near the nucleotide-binding domain of RAGA (Extended Data Fig. 9c), raising the question of whether the interaction we resolved

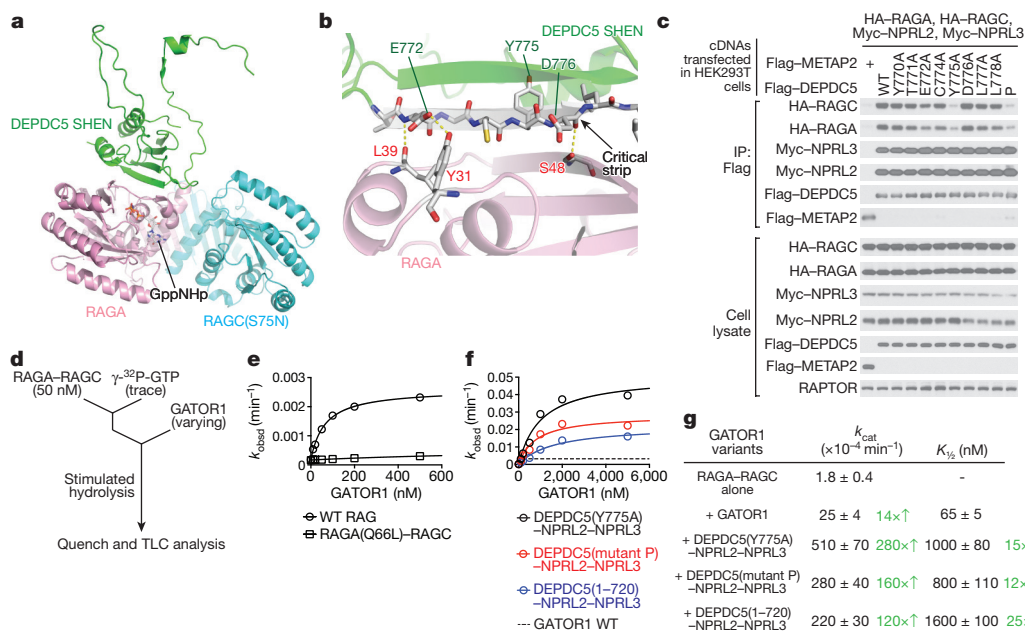


Figure 5 | The DEPDC5–Rag GTPases interaction represents an inhibitory state for GATOR1. **a**, Architecture of the SHEN domain–Rag GTPases interactions. **b**, The critical strip on DEPDC5 mediates the interaction with RAGA. Three pairs of hydrogen bonds are shown by yellow dashed lines. **c**, Point mutations in the critical strip of DEPDC5 impair binding of the Rag GTPase heterodimer to GATOR1. **d**, Single turnover stimulation assay. Scheme for single turnover GTP hydrolysis assay to determine the stimulatory effect of GATOR1 on the

Rag GTPases. **e**, **f**, Single turnover GTP hydrolysis stimulated by GATOR1 (**e**) or GATOR1 mutants (**f**). Dose-dependent GAP activity of wild-type GATOR1 (**e**) and variants that are defective in Rag GTPase binding (**f**). Representative datasets are shown here, and the statistics are summarized in **g**. **g**, Summary of kinetic parameters (single turnover, 25°C) for the GAP activity shown in **e** and **f**. Mean \pm s.d. of two to three independent experiments is reported. For gel source data, see Supplementary Figure 1.

here is the one responsible for stimulating GTP hydrolysis. To test this possibility, we purified GATOR1 variants containing the DEPDC5 mutants deficient in Rag GTPases binding and tested their GAP activity using a single turnover assay (Fig. 5d). Notably, these GATOR1 variants have enhanced GAP kinetics compared to the wild-type complex (Fig. 5e–g). For example, compared to wild-type GATOR1, the variant containing the DEPDC5(Y775A) mutant has a 20- and 15-fold increase in k_{cat} and $K_{1/2}$, respectively (Fig. 5f, g), indicating that a weaker interaction (increased $K_{1/2}$) carries out the real GAP function (increased k_{cat}). To further confirm this result, we generated a truncated DEPDC5, consisting of only residues 1–720, that completely lacks the SHEN domain. This truncated version of DEPDC5 still forms a complex with NPRL2–NPRL3 (Extended Data Fig. 9d), and has elevated hydrolysis kinetics similar to those of the GATOR1 variant containing DEPDC5(Y775A) (Fig. 5f, g).

To further validate this result, we designed a multiple turnover GTP hydrolysis assay (Extended Data Fig. 9e–h), in which the excess amount of Rag GTPases should have the opportunity to occupy the two binding modes simultaneously. Wild-type GATOR1 displayed a biphasic behaviour in its reaction kinetics: at lower concentrations of the Rag GTPases, the hydrolysis rate exhibited a transient plateau (inset in Extended Data Fig. 9f). At higher concentrations of the Rag GTPases, however, we observed additional stimulation, probably because the increased concentration of the Rag GTPases promoted a weaker interaction with a higher GAP activity (Extended Data Fig. 9f). Importantly, the initial phase was missing with the DEPDC5(Y775A) mutant (Extended Data Fig. 9g). These results suggest that the DEPDC5–RAGA contact detected in our structure does not execute the GAP activity of GATOR1, which must therefore be performed by an alternative interaction.

Two binding modes between GATOR1 and Rag GTPases

Based on the above data, we generated DEPDC5 in the absence of the NPRL2–NPRL3 heterodimer, and NPRL2–NPRL3 in the absence of DEPDC5, and then tested the capacity of each to GAP RAGA (Extended Data Fig. 9d). DEPDC5 had no activity, but NPRL2–NPRL3 robustly stimulated GTP hydrolysis by RAGA (Fig. 6a, b). Compared to intact GATOR1, a much higher concentration of NPRL2–NPRL3 was required to stimulate RAGA GTP hydrolysis, indicating that the absence of DEPDC5 substantially reduces the binding affinity between the Rag GTPases and NPRL2–NPRL3 (Fig. 6a, b). Moreover, excess NPRL2–NPRL3 stimulates hydrolysis even in the presence of wild-type GATOR1, suggesting that DEPDC5 prevents the NPRL2–NPRL3 within GATOR1 from accessing RAGA (Extended Data Fig. 9i, j). These results further support the idea that a weaker interaction—different from the one we that we observed—carries out the GAP function.

To independently confirm the binding between NPRL2–NPRL3 and the Rag GTPases, we performed a co-immunoprecipitation assay in cells. In cells lacking DEPDC5, the NPRL2–NPRL3 heterodimer co-immunoprecipitates the Rag GTPases (Fig. 6c). The interaction was enhanced by the presence of the RAGA(Q66L) mutant that prevents GTP hydrolysis (Fig. 6c, RAGA•GTP form), as well as by the presence of DEPDC5 that permits formation of the inhibitory binding mode (Extended Data Fig. 10a).

We further reasoned that if NPRL2–NPRL3 is the unit that acts as a GAP and is the receiver for amino acid signals (Extended Data Fig. 7), amino acid availability should regulate the interaction between NPRL2–NPRL3 and the Rag GTPases. To test this hypothesis directly, we pulled down NPRL2–NPRL3 in cells lacking DEPDC5, and probed for the Rag GTPases that co-immunoprecipitate with NPRL2–NPRL3 in the presence or absence of amino acids. Higher amounts of the Rag GTPases associated with NPRL2–NPRL3 in nutrient-deprived conditions (Extended Data Fig. 10b), but not in cells lacking GATOR2, which probably conveys amino acid signals to GATOR1 (Extended Data Fig. 10c). These results suggest that amino acid signals are transmitted

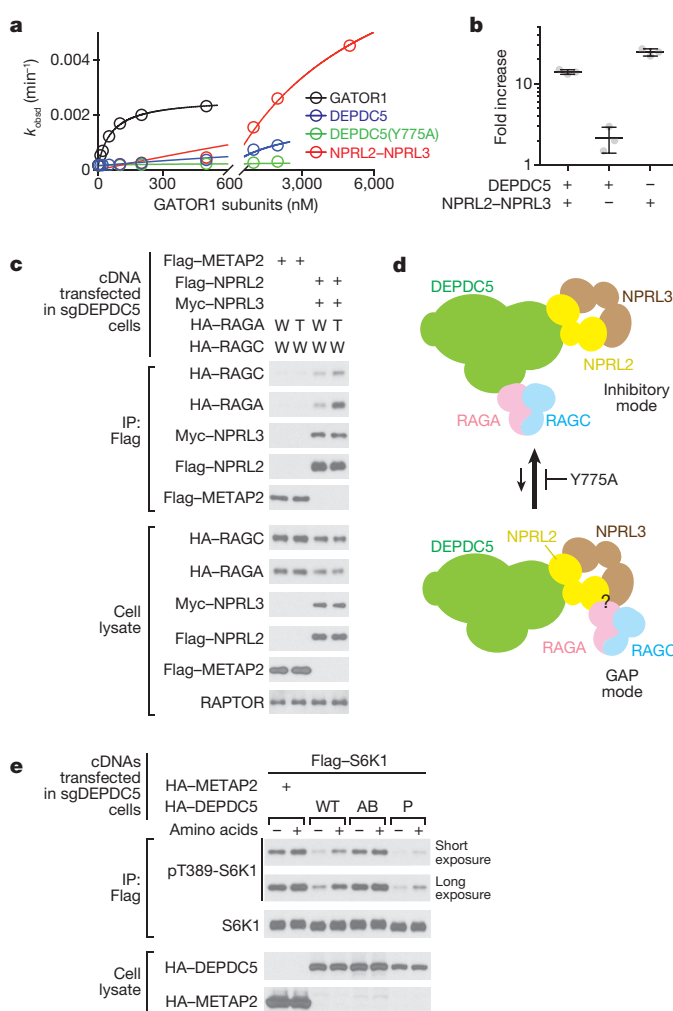


Figure 6 | A two-state model of the GATOR1 function. **a**, **b**, Single turnover GTP hydrolysis gives dose-dependent GAP activity of GATOR1 subunits DEPDC5 and NPRL2–NPRL3 (**a**), with quantification of rate enhancement by GATOR1 subunits (**b**). A representative dataset is shown in **a**, and the statistics are summarized in **b**. Mean \pm s.d. of three independent experiments is shown. **c**, Interaction between NPRL2–NPRL3 and the Rag GTPases in the absence of DEPDC5. W, wild-type RAGA or RAGC; T, RAGA(Q66L) mutant. **d**, A two-state model showing the equilibrium between GATOR1 and the Rag GTPases. Both the inhibitory mode and the GAP mode are required for regulating mTORC1 activity. **e**, Expression of a DEPDC5 mutant that is defective in Rag GTPases binding further suppresses mTORC1 activity in DEPDC5-null cells. AB, mutant AB. Data in **c** and **e** are representative of two independent experiments. For gel source data, see Supplementary Figure 1.

through GATOR2 to NPRL2–NPRL3 to directly regulate the Rag GTPases.

These results led us to conclude that at least two interaction modes must exist between the Rag GTPases and GATOR1 (Fig. 6d): an inhibitory mode characterized by a strong binding affinity between the Rag GTPases and the DEPDC5 SHEN domain, but a low GAP activity; and an alternative ‘GAP mode’ with the opposite characteristics. This proposal raised the question of the biological relevance of the inhibitory mode captured by our structure, as no similar behaviour has been previously observed for a GAP. To probe this question, we tested the effects on mTORC1 signalling of expressing DEPDC5 mutants deficient in Rag GTPases binding. We reasoned that if, as detected *in vitro* (Fig. 5f, g), the inhibitory mode suppresses the GAP activity of GATOR1 in cells, we should observe lower mTORC1 signalling (that is, enhanced GAP activity) when we eliminate it. This is indeed the case: in cells expressing mutant P of DEPDC5, mTORC1 signalling was more suppressed

than in those expressing wild-type DEPDC5 even under nutrient-rich conditions (Fig. 6e). Moreover, this increased degree of inhibition requires NPRL2–NPRL3, as we saw no difference between mutant P and wild-type DEPDC5 in cells lacking NPRL2 (Extended Data Fig. 10d), which further supports the notion that the NPRL2–NPRL3 heterodimer carries out the GAP activity of GATOR1. We therefore conclude that the inhibitory mode between GATOR1 and the Rag GTPases operates within cells and serves to prevent GATOR1 hyperactivation to maintain the proper response of mTORC1 to nutrients.

Summary

In this study we present cryo-EM structures for GATOR1 and the GATOR1–Rag GTPases complex. Our work suggests that at least two binding modes exist between GATOR1 and the Rag GTPases and that both are required for mTORC1 signalling to respond normally to nutrients. The inhibitory mode we have identified distinguishes GATOR1 from canonical GAPs and represents an unforeseen mechanism for how cells suppress mTORC1 activity under nutrient-deficient conditions.

Data availability Atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank (PDB) with accession numbers 6CET for GATOR1 and 6CES for GATOR1–Rag GTPases. Electron density maps have been deposited in Electron Microscopy Data Bank with accession numbers EMD-7465 for GATOR1 and EMD-7464 for GATOR1–Rag GTPases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 October 2017; accepted 16 February 2018.

Published online 28 March 2018.

- Efeyan, A., Comb, W. C. & Sabatini, D. M. Nutrient-sensing mechanisms and pathways. *Nature* **517**, 302–310 (2015).
- Shaw, R. J. & Cantley, L. C. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature* **441**, 424–430 (2006).
- Jewell, J. L., Russell, R. C. & Guan, K.-L. Amino acid signalling upstream of mTOR. *Nat. Rev. Mol. Cell Biol.* **14**, 133–139 (2013).
- González, A. & Hall, M. N. Nutrient sensing and TOR signaling in yeast and mammals. *EMBO J.* **36**, 397–408 (2017).
- Saxton, R. A. & Sabatini, D. M. mTOR signaling in growth, metabolism, and disease. *Cell* **168**, 960–976 (2017).
- Schürmann, A., Brauers, A., Massmann, S., Becker, W. & Joost, H. G. Cloning of a novel family of mammalian GTP-binding proteins (RagA, RagBs, RagB1) with remote similarity to the Ras-related GTPases. *J. Biol. Chem.* **270**, 28982–28988 (1995).
- Hirose, E., Nakashima, N., Sekiguchi, T. & Nishimoto, T. RagA is a functional homologue of *S. cerevisiae* Gtr1p involved in the Ran/Gsp1-GTPase pathway. *J. Cell Sci.* **111**, 11–21 (1998).
- Sekiguchi, T., Hirose, E., Nakashima, N., Li, M. & Nishimoto, T. Novel G proteins, Rag C and Rag D, interact with GTP-binding proteins, Rag A and Rag B. *J. Biol. Chem.* **276**, 7246–7257 (2001).
- Nakashima, N., Noguchi, E. & Nishimoto, T. *Saccharomyces cerevisiae* putative G protein, Gtr1p, which forms complexes with itself and a novel protein designated as Gtr2p, negatively regulates the Ran/Gsp1p G protein cycle through Gtr2p. *Genetics* **152**, 853–867 (1999).
- Sancak, Y. et al. The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320**, 1496–1501 (2008).
- Inoki, K., Li, Y., Xu, T. & Guan, K.-L. Rheb GTPase is a direct target of TSC2 GAP activity and regulates mTOR signaling. *Genes Dev.* **17**, 1829–1834 (2003).
- Menon, S. et al. Spatial control of the TSC complex integrates insulin and nutrient regulation of mTORC1 at the lysosome. *Cell* **156**, 771–785 (2014).
- Li, Y., Inoki, K. & Guan, K.-L. Biochemical and functional characterizations of small GTPase Rheb and TSC2 GAP activity. *Mol. Cell Biol.* **24**, 7965–7975 (2004).
- Saito, K., Araki, Y., Kontani, K., Nishina, H. & Katada, T. Novel role of the small GTPase Rheb: its implication in endocytic pathway independent of the activation of mammalian target of rapamycin. *J. Biochem.* **137**, 423–430 (2005).
- Saucedo, L. J. et al. Rheb promotes cell growth as a component of the insulin/TOR signalling network. *Nat. Cell Biol.* **5**, 566–571 (2003).
- Stocker, H. et al. Rheb is an essential regulator of S6K in controlling cell growth in *Drosophila*. *Nat. Cell Biol.* **5**, 559–566 (2003).
- Shen, K., Choe, A. & Sabatini, D. M. Intersubunit crosstalk in the Rag GTPase heterodimer enables mTORC1 to respond rapidly to amino acid availability. *Mol. Cell* **68**, 552–565.e8 (2017).
- Panchaud, N., Péli-Gulli, M.-P. & De Virgilio, C. Amino acid deprivation inhibits TORC1 through a GTPase-activating protein complex for the Rag family GTPase Gtr1. *Sci. Signal.* **6**, ra42 (2013).
- Bar-Peled, L. et al. A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. *Science* **340**, 1100–1106 (2013).
- Petit, C. S., Roczniak-Ferguson, A. & Ferguson, S. M. Recruitment of folliculin to lysosomes supports the amino acid-dependent activation of Rag GTPases. *J. Cell Biol.* **202**, 1107–1122 (2013).
- Tsun, Z.-Y. et al. The folliculin tumor suppressor is a GAP for the RagC/D GTPases that signal amino acid levels to mTORC1. *Mol. Cell* **52**, 495–505 (2013).
- Dibbens, L. M. et al. Mutations in DEPDC5 cause familial focal epilepsy with variable foci. *Nat. Genet.* **45**, 546–551 (2013).
- Ishida, S. et al. Mutations of DEPDC5 cause autosomal dominant focal epilepsies. *Nat. Genet.* **45**, 552–555 (2013).
- Wu, X. & Tu, B. P. Selective regulation of autophagy by the Iml1–Npr2–Npr3 complex in the absence of nitrogen starvation. *Mol. Biol. Cell* **22**, 4124–4133 (2011).
- Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40 (2008).
- Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).
- Levine, T. P. et al. Discovery of new Longin and Roadblock domains that form platforms for small GTPases in Regulator and TRAPP-II. *Small GTPases* **4**, 62–69 (2013).
- Gong, R. et al. Crystal structure of the Gtr1p–Gtr2p complex reveals new insights into the amino acid-induced TORC1 activation. *Genes Dev.* **25**, 1668–1673 (2011).
- Jeong, J.-H. et al. Crystal structure of the Gtr1p(GTP)–Gtr2p(GDP) protein complex reveals large structural rearrangements triggered by GTP-to-GDP conversion. *J. Biol. Chem.* **287**, 29648–29653 (2012).
- Gibrat, J. F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385 (1996).
- Shiozawa, K. et al. Structure of the N-terminal domain of PEX1 AAA-ATPase. Characterization of a putative adaptor-binding domain. *J. Biol. Chem.* **279**, 50060–50068 (2004).
- Ingelman, M., Bianchi, V. & Eklund, H. The three-dimensional structure of flavodoxin reductase from *Escherichia coli* at 1.7 Å resolution. *J. Mol. Biol.* **268**, 147–157 (1997).
- Qu, A. & Leahy, D. J. Crystal structure of the I-domain from the CD11a/CD18 (LFA-1, alpha L beta 2) integrin. *Proc. Natl Acad. Sci. USA* **92**, 10277–10281 (1995).
- Algre, R. et al. Molecular architecture and function of the SEA complex, a modulator of the TORC1 pathway. *Mol. Cell Proteomics* **13**, 2855–2870 (2014).
- Kurzbaue, R. et al. Crystal structure of the p14/MP1 scaffolding complex: how a twin couple attaches mitogen-activated protein kinase signaling to late endosomes. *Proc. Natl Acad. Sci. USA* **101**, 10984–10989 (2004).
- Wolfson, R. L. et al. KICSTOR recruits GATOR1 to the lysosome and is necessary for nutrients to regulate mTORC1. *Nature* **543**, 438–442 (2017).
- Gu, X. et al. SAMTOR is an S-adenosylmethionine sensor for the mTORC1 pathway. *Science* **358**, 813–818 (2017).
- Scheffzek, K. et al. The Ras–RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* **277**, 333–338 (1997).
- Daumke, O., Weyand, M., Chakrabarti, P. P., Vetter, I. R. & Wittinghofer, A. The GTPase-activating protein Rap1GAP uses a catalytic asparagine. *Nature* **429**, 197–201 (2004).

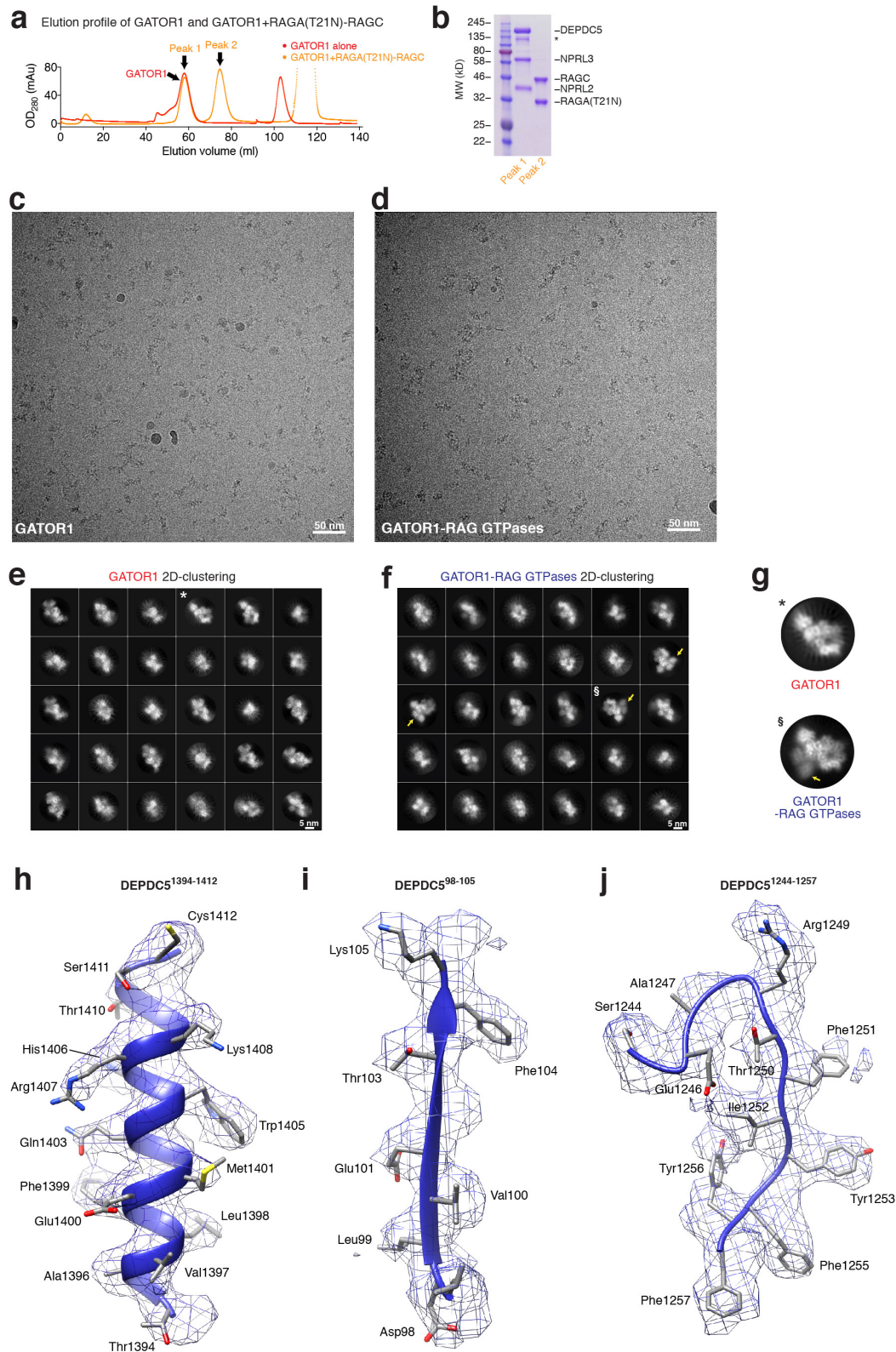
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all members of the Sabatini laboratory and T. Schwartz for insights; P. Abeyaratne, N. Grigorieff, R. Grant and C. Drennan for technical support; R. Saxton, M. Pacold and S. Shan for critical reading of the manuscript. This work was supported by grants from the NIH (R01 CA103866, R01 CA129105 and R37 AI047389) and Department of Defense (W81XWH-15-1-0230) to D.M.S., fellowship support from NSF (2016197106) to K.J.C. and from the Life Sciences Research Foundation to K.S., where he is a Pfizer Fellow. R.K.H., C.H. and Z.Y. were supported by the Howard Hughes Medical Institute. D.M.S. is an investigator of the Howard Hughes Medical Institute.

Author Contributions K.S. and D.M.S. initiated the project. K.S. purified the proteins and performed the biochemical characterization with input from K.J.C., M.L.V., L.C., A.B., and A.C. R.K.H., C.H. and Z.Y. determined the electron microscopy density maps for GATOR1 and GATOR1–Rag GTPases. K.S. and E.J.B. built the structural model. K.S., R.K.H., E.J.B., Z.Y. and D.M.S. wrote and edited the manuscript.

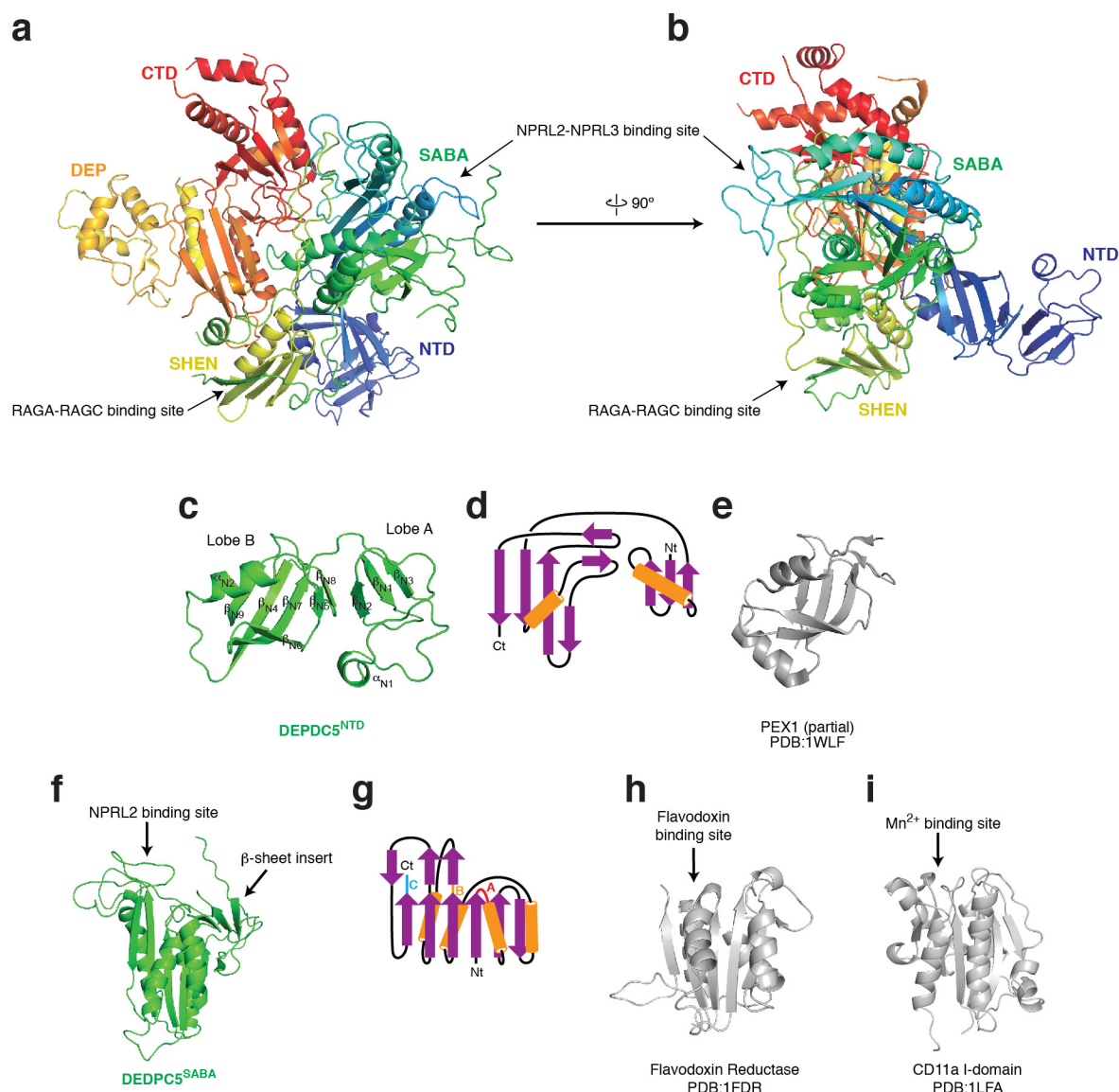
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Z.Y. (yuz@janelia.hhmi.org) or D.M.S. (sabatini@wi.mit.edu).

Reviewer Information Nature thanks D. Barford, K. Inoki and the other anonymous reviewer(s) for their contribution to the peer review of this work.



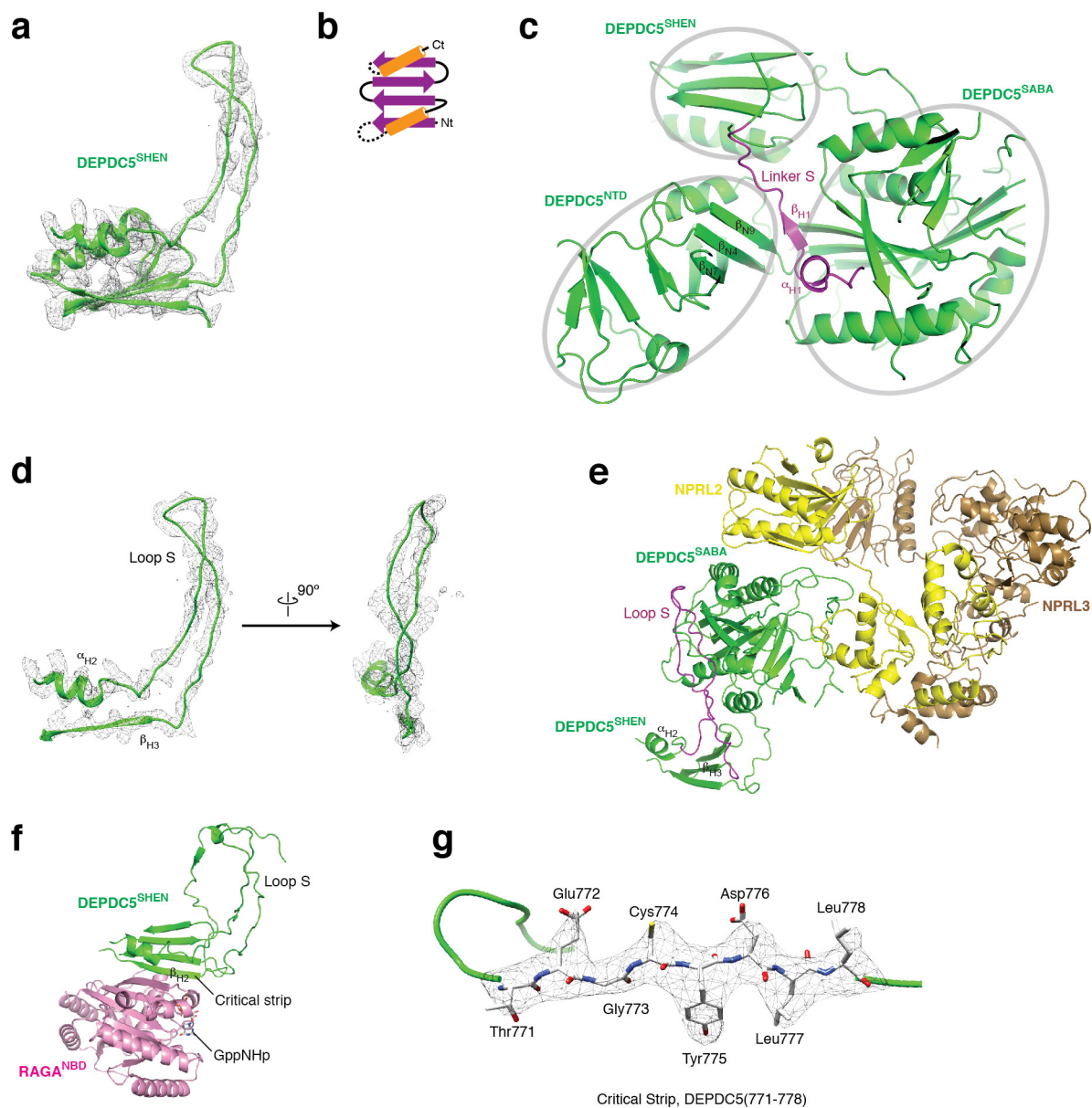
Extended Data Figure 1 | Structural determination and model building for the GATOR1 and GATOR1-Rag GTPases complex. **a**, Gel filtration profiles for GATOR1 (red line) and GATOR1 + RAGA(T21N)-RAGC (orange line). The peak position for GATOR1 does not shift upon incubation with RAGA(T21N)-RAGC, which suggests that there is no direct binding between the two complexes. **b**, Coomassie blue stained SDS-PAGE analysis of the two peaks on the GATOR1 + RAGA(T21N)-RAGC elution profile. No co-elution is observed. Asterisk denotes a non-specific band that co-purifies with GATOR1. **c**, **d**, Raw cryo-EM images for GATOR1 (**c**) and the GATOR1-Rag GTPases complex (**d**). Discrete particles were clearly visualized under the microscope. Scale bars, 50 nm.

e, **f**, 2D clustering of GATOR1 (**e**) and GATOR1-Rag GTPases (**f**). Yellow arrows in **f** point to the extra electron microscopy densities in comparison to **e**. * and § mark the particles shown in **g**. Scale bars, 5 nm. **g**, Direct comparison of particles from 2D clustering of GATOR1 and GATOR1-Rag GTPases. Extra electron microscopy densities for the Rag GTPases can be directly observed. **h**–**j**, Extracted regions from the electron microscopy density maps and the fitted structures for α -helical (**h**), β -strand (**i**) and loop (**j**) regions of DEPDC5. Secondary structures and bulky side chains can be unambiguously resolved at the current resolution. Data in **a** and **b** are representative of two independent experiments.



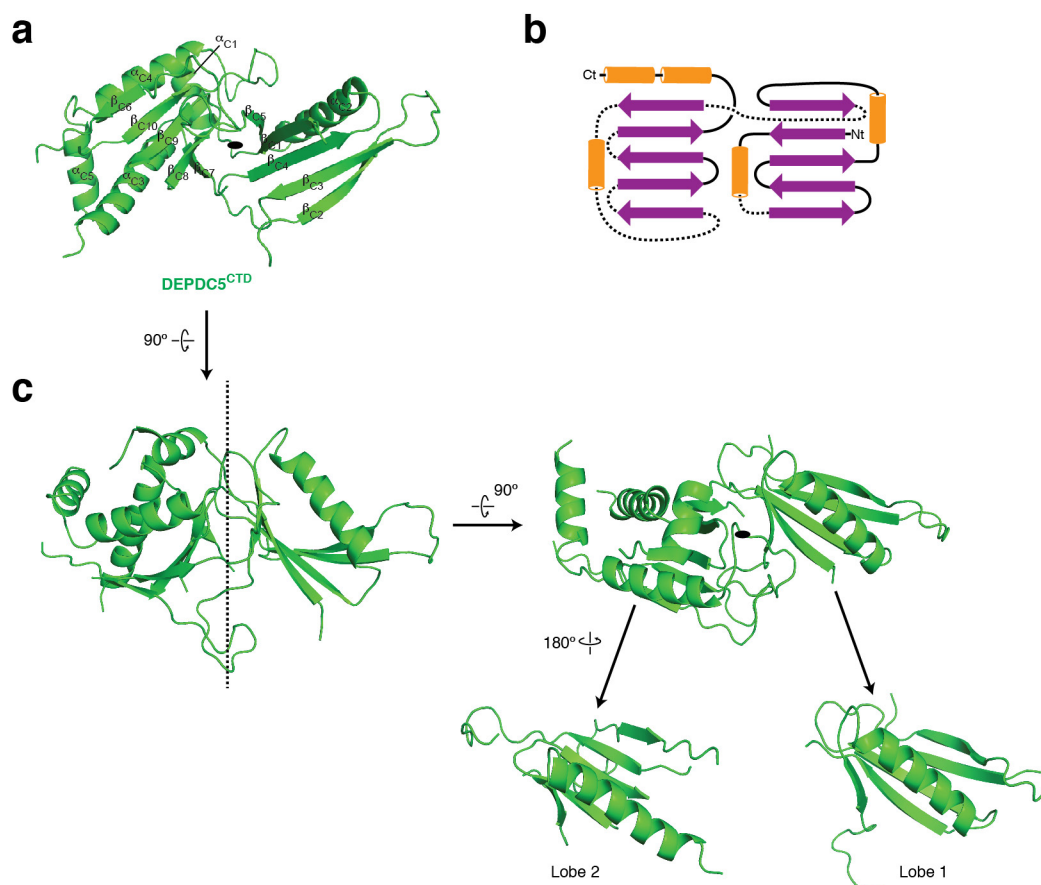
Extended Data Figure 2 | Architecture of DEPDC5. **a, b**, Two views of DEPDC5. The protein backbone is depicted in rainbow colours from the N terminus (blue) to the C terminus (red). Binding sites for NPRL2–NPRL3 and the Rag GTPases are marked. **c, d**, Structural model (**c**) and topological diagram (**d**) for DEPDC5 NTD. **e**, Lobe B of DEPDC5 NTD shares structural similarity to the NTD of the PEX1 AAA-ATPase. **f–i**, The SABA domain of DEPDC5 (**f**) shares topological similarity (**g**)

to flavodoxin reductase (**h**) and CD11a I domain (**i**), which all contain ligand-binding sites (indicated by arrows). The SABA domain contains a β -sheet insertion formed by three strands. The three loops in the SABA domain of DEPDC5 that mediate the DEPDC5–NPRL2 interaction are coloured in red (loop A), orange (loop B), and blue (loop C), respectively, on the topological diagram. Nt, N terminus; Ct, C terminus.

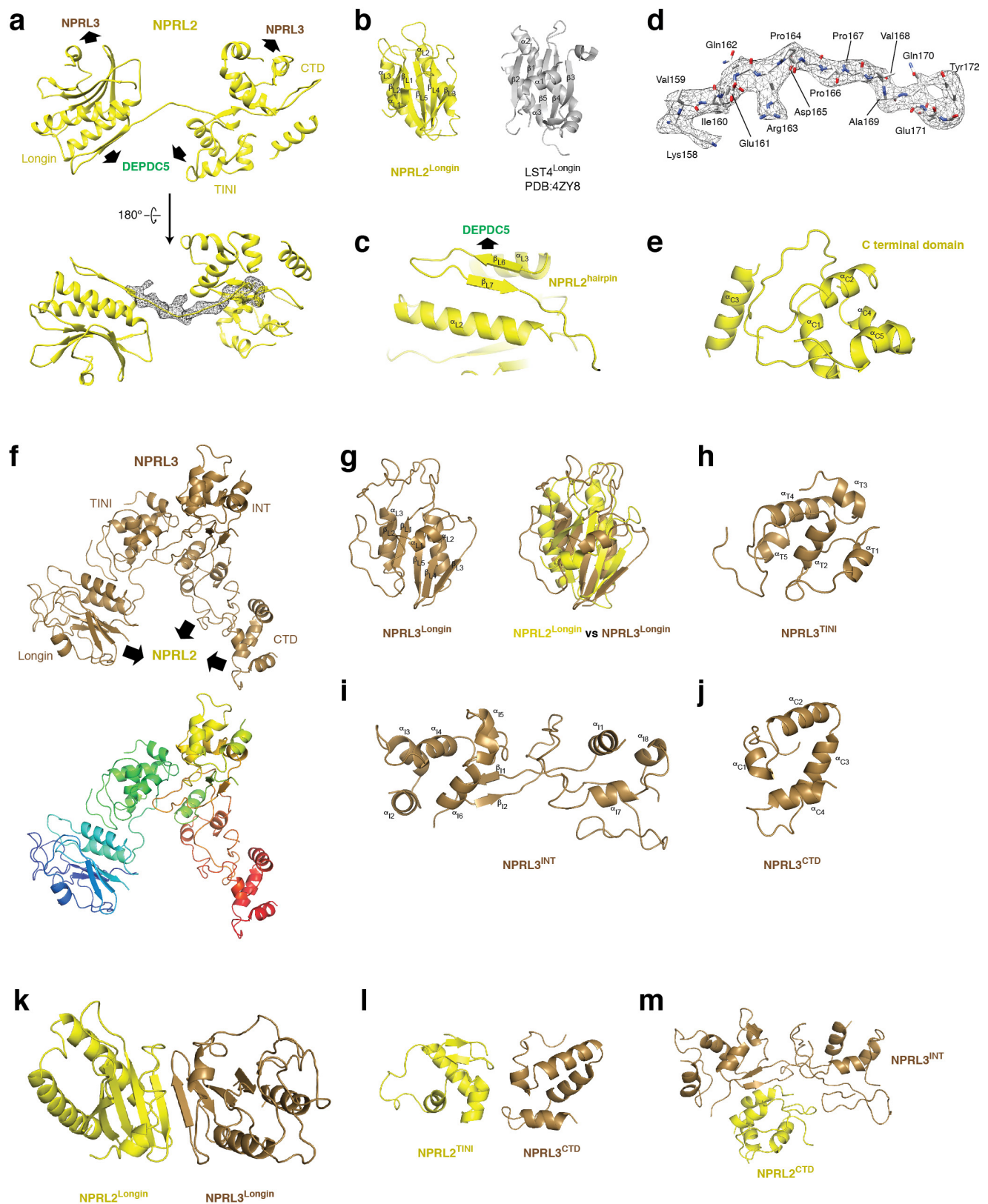


Extended Data Figure 3 | Architecture of the SHEN domain of DEPDC5. **a**, Electron microscopy density map and structural model for the SHEN domain. **b**, Topological diagram for the SHEN domain. **c**, β_{H1} on linker S forms a continuous sheet with the β -strands on lobe B of the NTD, and positions itself between the NTD and the SABA domain. **d**, Electron microscopy density map and structural model for loop S.

e, Loop S (purple) mediates interdomain contact with the SABA domain of DEPDC5, close to where the NPRL2–NPRL3 dimer binds to DEPDC5. **f**, β_{H2} (which we named the critical strip) of the SHEN domain directly contacts RAGA (pink). **g**, Electron microscopy density map and the atomic model for the critical strip.



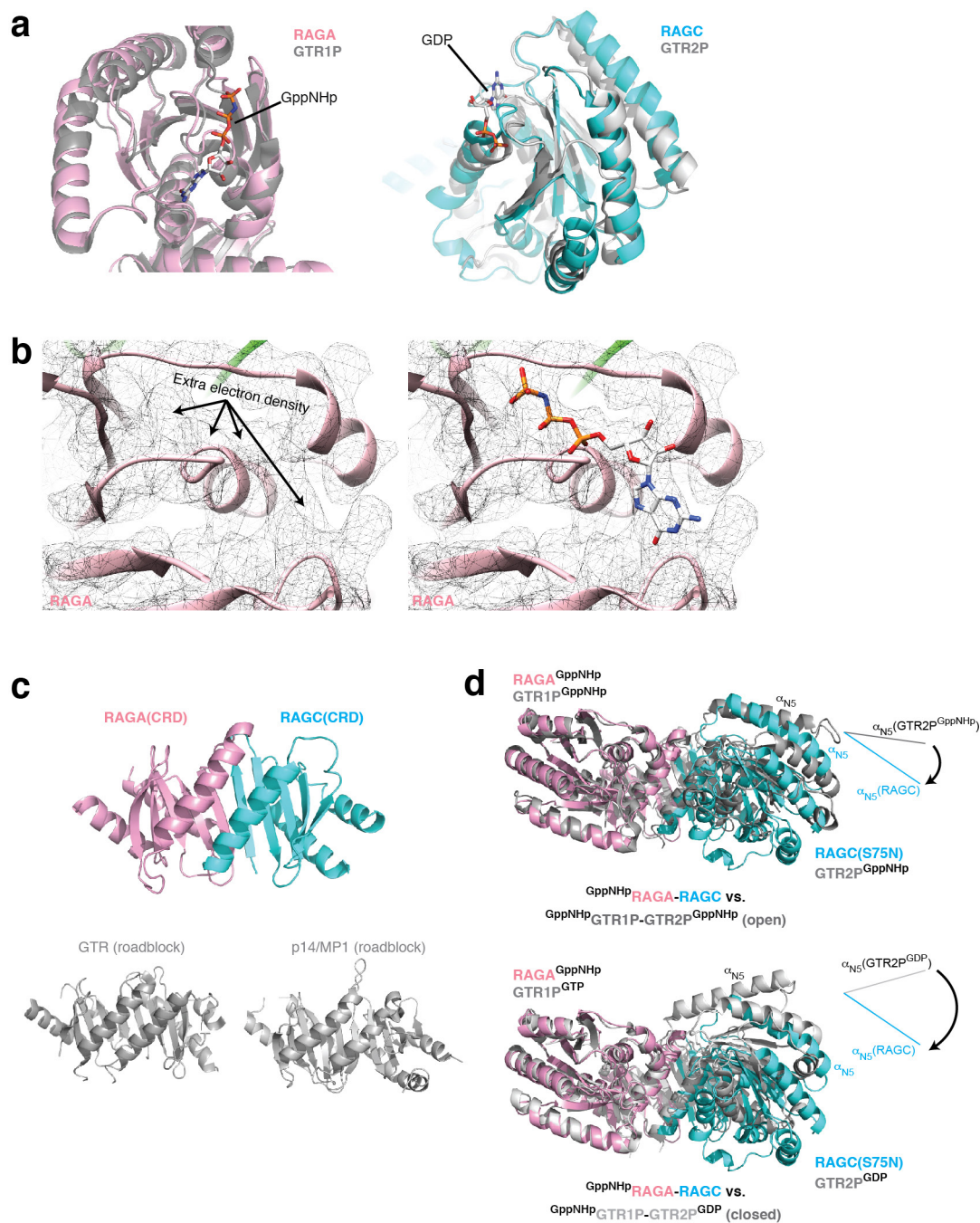
Extended Data Figure 4 | Architecture of the CTD of DEPDC5. **a, b**, Structure (**a**) and topological diagram (**b**) for the CTD of DEPDC5. **c**, The CTD of DEPDC5 shows a pseudo two-fold rotational symmetry. Two lobes with similar folds can be seen.



Extended Data Figure 5 | Architecture of NPRL2 and NPRL3.

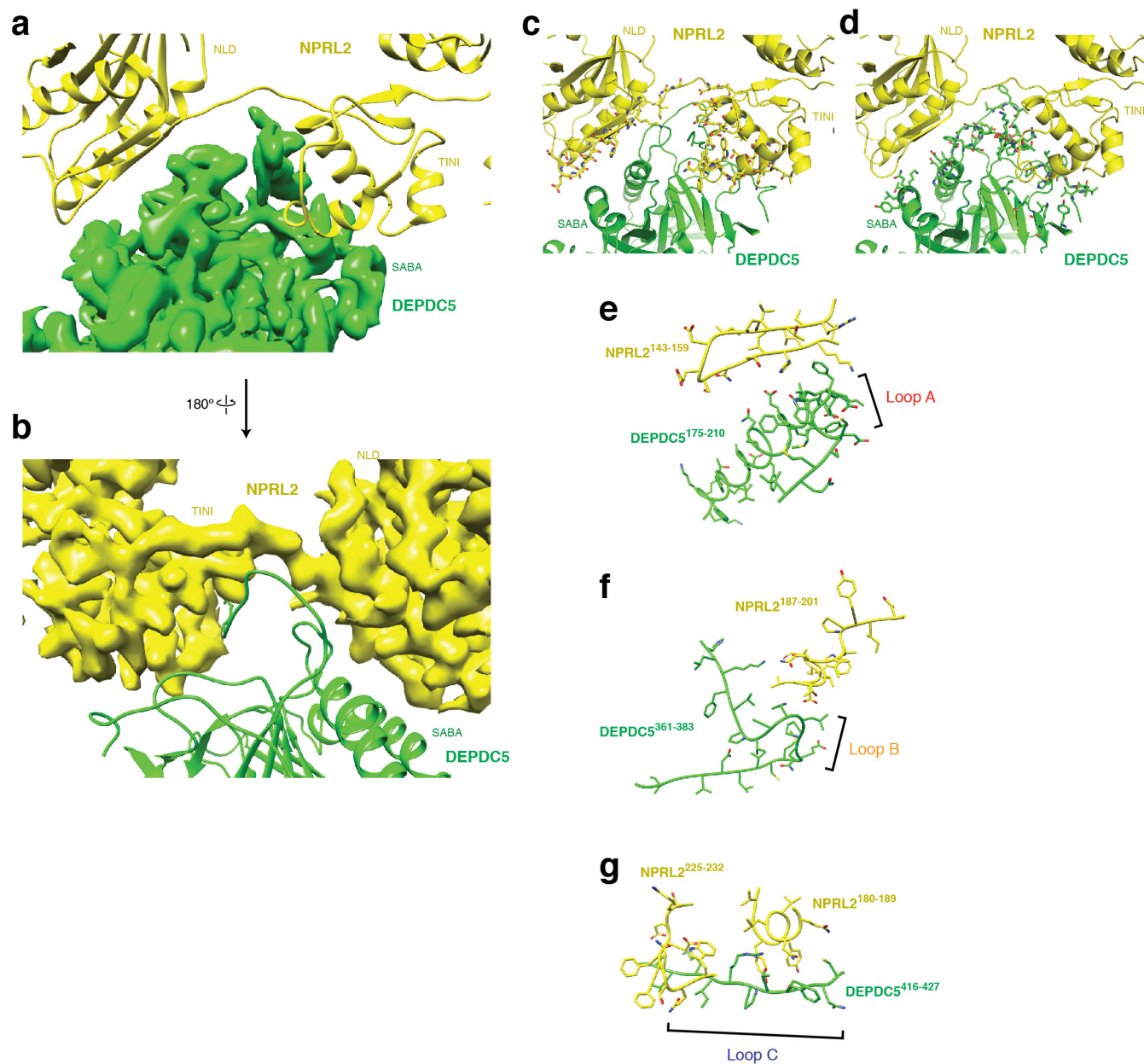
a, Structural model of NPRL2. Contact surfaces with DEPDC5 and NPRL3 are indicated by arrows. A long linker connects the longin domain and the TINI domain, with electron microscopy density shown as mesh. The atomic model for this linker is shown in **d**. **b**, Longin domain of NPRL2. A standard longin domain from LST4 is shown for comparison. **c**, A strand-turn-strand motif (hairpin) is attached to the longin domain of NPRL2, which mediates partial interaction with DEPDC5. **d**, Electron microscopy density map and atomic model for the linker connecting the longin domain and the TINI domain (see the electron microscopy density in **a**).

e, Structural model for the CTD of NPRL2. **f**, Structural model for NPRL3. Contact surfaces with NPRL2 are indicated by arrows. **g**, Longin domain of NPRL3 and its overlap with the longin domain of NPRL2. **h**, Structural model for the TINI domain of NPRL3 that connects its longin domain with the CTDs. **i**, Structural model for the intermediary (INT) domain of NPRL3. **j**, Structural model for the CTD of NPRL3. **k–m**, Interactions between NPRL2 and NPRL3. Three contact surfaces were identified that mediate the interactions between NPRL2 and NPRL3: the longin domains of NPRL2 and NPRL3 (**k**), the TINI domain of NPRL2 and CTD of NPRL3 (**l**), and the CTD of NPRL2 and the INT domain of NPRL3 (**m**).



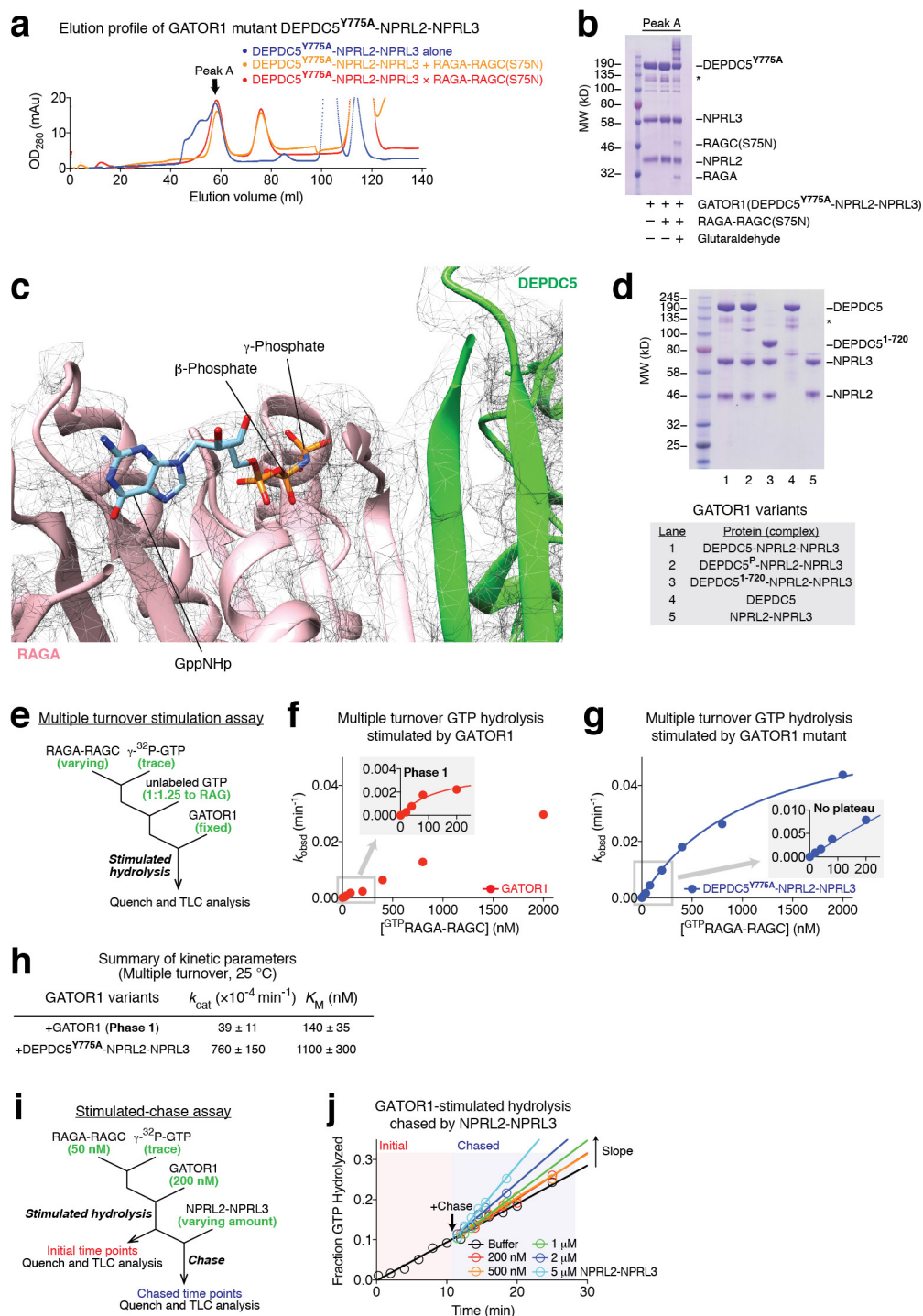
Extended Data Figure 6 | Architecture of the Rag GTPase heterodimer. **a**, NBDs of RAGA (pink) and RAGC (cyan) overlap with those of Gtr1p and Gtr2p (grey). **b**, Extra electron microscopy density can be observed in the nucleotide-binding pocket of RAGA, into which GppNHp can be fitted. **c**, The CRD of RAGA and RAGC tightly dimerize with one another. The dimerized roadblock domains from Gtr1p-Gtr2p and p14-MP1

are shown for comparison. **d**, Global conformation of the Rag GTPase heterodimer in comparison to the two crystal structures of Gtr1p-Gtr2p. RAGA and Gtr1p are aligned. Rotational movement of the NBD of RAGC is illustrated, and compared with the NBD of Gtr2p in the direction of α_{N5} . The NBDs of the Rag GTPases rotate further away from one another, even when compared with the open conformation of Gtr1p-Gtr2p (top).



Extended Data Figure 8 | Interactions between DEPDC5 and NPRL2.
a, b, Large contact surfaces between DEPDC5 (green) and NPRL2 (yellow) are observed from the electron microscopy density map and structural models. **c, d**, Surface residues on NPRL2 (**c**) and DEPDC5 (**d**)

participate in mediating interactions between the two proteins, identified by 'InterfaceResidue' script in Pymol. **e–g**, Loops A (**e**), B (**f**) and C (**g**) on DEPDC5 directly contact NPRL2.

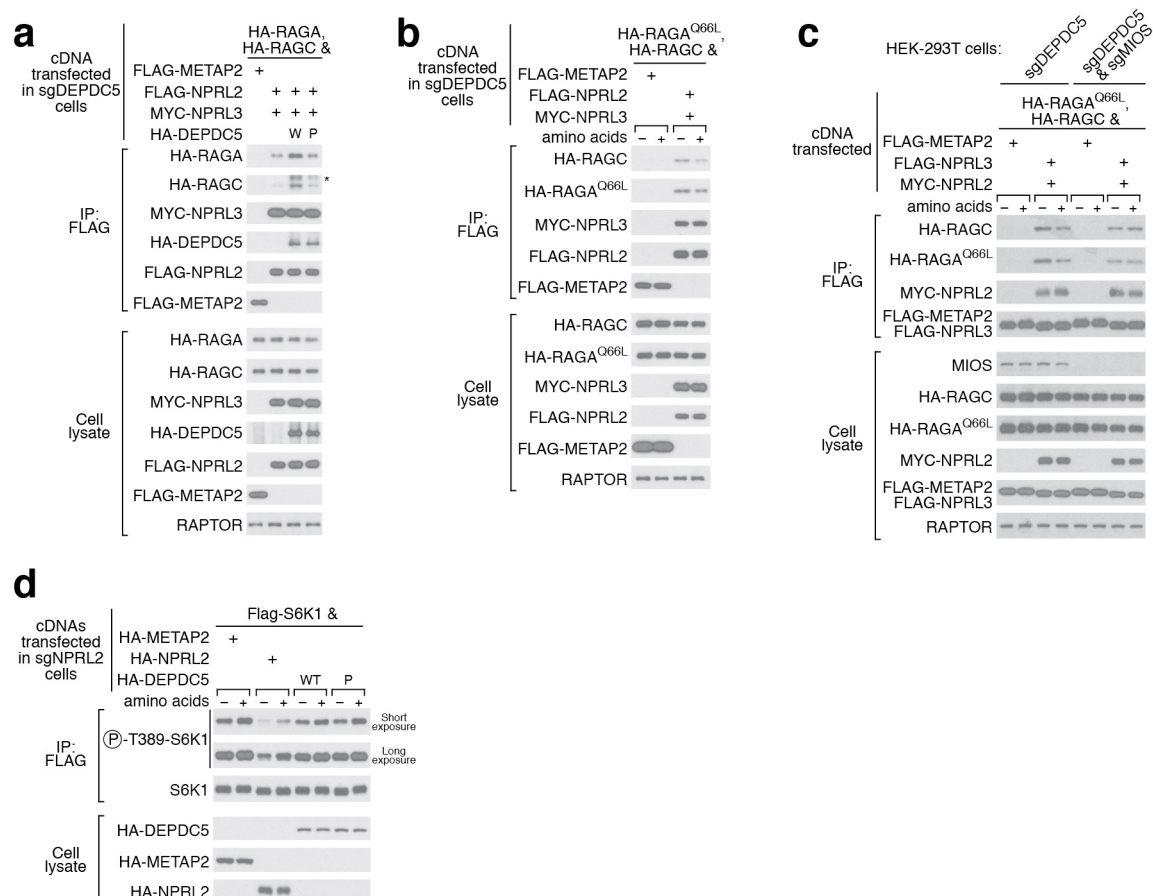


Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | *In vitro* characterization of the GAP

mechanism of GATOR1. **a**, Gel filtration profiles for DEPDC5(Y775A)–NPRL2–NPRL3 (blue line) and DEPDC5(Y775A)–NPRL2–NPRL3 + RAGA–RAGC(S75N) in the absence (orange line) or presence (red line) of the crosslinker glutaraldehyde. Peak A denotes the species eluted at the large molecular weight region. **b**, Coomassie blue stained SDS–PAGE analysis of peak A. Direct binding is only observed in the presence of glutaraldehyde. Asterisk denotes a non-specific band that co-purifies with GATOR1. **c**, No extra electron microscopy density can be observed near the NBD of RAGA. **d**, GATOR1 variants visualized by SDS–PAGE followed by Coomassie blue staining. Asterisk denotes a non-specific band that co-purifies with GATOR1. **e**, Scheme for measuring stimulated GTP hydrolysis by GATOR1 in a multiple turnover setup. An excess amount of Rag GTPases singly loaded with GTP was incubated with fixed amount of GATOR1. The hydrolysis reaction was traced and quantified. **f**, Stimulated GTP hydrolysis by wild-type GATOR1 shows a biphasic behaviour in reaction kinetics. As an increasing amount of the Rag GTPases was included in the reaction, a small plateau of observed rate constant (k_{obsd}) was first observed at a lower concentration (inset). Such biphasic behaviour indicates that two binding modes exist in the wild-type GATOR1: one with higher affinity to the Rag GTPases but lower GAP activity, the other with lower affinity but higher GAP activity.

A representative dataset is shown in this panel, and the statistics are summarized below. **g**, Stimulated GTP hydrolysis by a GATOR1 mutant that is defective in Rag GTPases binding eliminates the initial phase. DEPDC5(Y775A)–NPRL2–NPRL3 is defective in stable Rag GTPases binding because it lacks the docking site (an intact critical strip) for the Rag GTPases. Consequentially, the inhibitory mode diminishes (inset), leaving a single phase corresponding to the GAP mode in reaction kinetics. A representative dataset is shown, and the statistics are summarized below. **h**, Summary of kinetic parameters for the multiple turnover GAP activity shown in **f** and **g**. Mean \pm s.d. of two to three independent experiments is reported. **i**, A stimulated-chase assay to characterize the inhibition mechanism of DEPDC5. Wild-type GATOR1 was first added to bind the Rag GTPases with its inhibitory mode. Extra NPRL2–NPRL3 was then included in the reaction as a chase. We reasoned that if DEPDC5 sequesters the NBD of RAGA, no further stimulation should be observed; if DEPDC5 simply prevents NPRL2–NPRL3 from accessing the NBD of RAGA, we should observe additional stimulation because there is no DEPDC5 to inhibit the extra NPRL2–NPRL3. **j**, Further stimulation is observed in the presence of additional NPRL2–NPRL3, as reflected by the faster hydrolysis rate (steeper slope), suggesting DEPDC5 inhibits NPRL2–NPRL3 *in cis*. Data in **a**, **b**, **d** and **j** are representative of two independent experiments.



Extended Data Figure 10 | *In vivo* characterization of the GAP mechanism of GATOR1. a, Interaction between NPRL2–NPRL3 and the Rag GTPases is enhanced by wild-type DEPDC5 but not mutant P, which is defective in binding to the Rag GTPases. W, wild-type DEPDC5; P, mutant P. Asterisk denotes a non-specific band. **b**, Amino acid availability regulates the interaction between NPRL2–NPRL3 and the Rag GTPases in cells lacking DEPDC5. Higher amount of Rag GTPases co-

immunoprecipitates with NPRL2–NPRL3 in the absence of amino acids. **c**, Loss of regulated interaction between NPRL2–NPRL3 and the Rag GTPases in cells lacking DEPDC5 and MIOS. No difference is observed when GATOR2, the receptor for amino acid signals, is knocked out. **d**, Expression of a *Dedpc5* mutant that is defective in Rag GTPases binding has no effect in NPRL2-null cells, in sharp contrast to the result in Fig. 6e. Data in **a–d** are representative of two independent experiments.

A density cusp of quiescent X-ray binaries in the central parsec of the Galaxy

Charles J. Hailey¹, Kaya Mori¹, Franz E. Bauer^{2,3,4}, Michael E. Berkowitz¹, Jaesub Hong⁵ & Benjamin J. Hord¹

The existence of a ‘density cusp’^{1,2}—a localized increase in number—of stellar-mass black holes near a supermassive black hole is a fundamental prediction of galactic stellar dynamics³. The best place to detect such a cusp is in the Galactic Centre, where the nearest supermassive black hole, Sagittarius A*, resides. As many as 20,000 black holes are predicted to settle into the central parsec of the Galaxy as a result of dynamical friction^{3–5}; however, so far no density cusp of black holes has been detected. Low-mass X-ray binary systems that contain a stellar-mass black hole are natural tracers of isolated black holes. Here we report observations of a dozen quiescent X-ray binaries in a density cusp within one parsec of Sagittarius A*. The lower-energy emission spectra that we observed in these binaries is distinct from the higher-energy spectra associated with the population of accreting white dwarfs that dominates the central eight parsecs of the Galaxy⁶. The properties of these X-ray binaries, in particular their spatial distribution and luminosity function, suggest the existence of hundreds of binary systems in the central parsec of the Galaxy and many more isolated black holes. We cannot rule out a contribution to the observed emission from a population (of up to about one-half the number of X-ray binaries) of rotationally powered, millisecond pulsars. The spatial distribution of the binary systems is a relic of their formation history, either in the stellar disk around Sagittarius A* (ref. 7) or through in-fall from globular clusters, and constrains the number density of sources in the modelling of gravitational waves from massive stellar remnants^{8,9}, such as neutron stars and black holes.

The Chandra X-ray Observatory has accumulated 1.4×10^6 s of observations of the Galactic Centre using the Advanced CCD Imaging Spectrometer I (ACIS-I) over the past 12 years. Owing to the high concentration of X-ray sources in the Galactic Centre (such as Sagittarius A* (hereafter Sgr A*) and IRS 13) and the emission from hot gas, we restrict our analysis to angular distances of more than 5'' from Sgr A*, which correspond to projected distances r of more than 0.2 pc assuming an 8-kpc distance to the Galactic Centre¹⁰. From these Chandra observations, 415 X-ray point sources at projected distances of $0.2 \text{ pc} < r < 3.8 \text{ pc}$ from Sgr A* were identified in the 2–8-keV energy band, and used for all analysis. Point sources were ignored if they were in diffuse regions such as filamentary structures or molecular clouds, were too close to very bright point sources or were previously known to have large X-ray outbursts. A net count (source minus background counts) limit of $C \geq 100$ (equivalent to a flux of $2 \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$) was used for the analysis. This count limit corresponds to a point source detection significance of 4σ . There are 92 sources with $C \geq 100$, 26 of which lie at $r < 1 \text{ pc}$. For spectral fitting, a limit of $C \geq 200$ was used in order to constrain spectral parameters effectively: 13 sources with $r < 1 \text{ pc}$ met this criterion. This is too few sources for population analysis, so we instead use a hardness ratio of X-ray colour, or slope, $\text{HR2} = (C_{\text{H}} - C_{\text{L}})/(C_{\text{H}} + C_{\text{L}})$, where C_{L} and C_{H} are, respectively, the net counts from the source in the 2–4-keV (low) and 4–8-keV (high)

energy bands. Previous Chandra analysis¹¹ and simulations reported here demonstrate that the thermal emission of magnetic cataclysmic variable stars—which have typical temperatures of $kT \approx 8\text{--}40 \text{ keV}$ —in the Chandra energy band of 0.5–8 keV is well modelled by a single-temperature, optically thin thermal plasma and an iron line complex at an energy of $E = 6\text{--}7 \text{ keV}$. By contrast, black-hole and neutron-star binaries and pulsars are universally characterized by non-thermal power-law emission with a flux described by $F(E) = kE^{-\Gamma}$ photons $\text{cm}^{-2} \text{ s}^{-1} \text{ keV}^{-1}$, where E is the photon energy in keV and $\Gamma \approx 1.5\text{--}2$ (ref. 12) is the photon index. The low- and high-energy bands for HR2 were chosen to maximally separate thermal and non-thermal sources. Thermal sources have HR2 values of more than 0.3, whereas non-thermal sources have HR2 values of less than 0.3.

In Fig. 1a, b we show the HR2 distributions for sources with $C \geq 100$ that lie in the circle $r < 1 \text{ pc}$ and in the annulus $1 \text{ pc} < r < 3.8 \text{ pc}$, respectively, centred on Sgr A*. A second source population clearly appears at $r < 1 \text{ pc}$ in the range $-0.1 < \text{HR2} < 0.3$. Integrating the HR2 distribution in the annular region for $\text{HR2} < 0.3$, we would expect at most one source with $\text{HR2} < 0.3$ for $r < 1 \text{ pc}$ if the sources in the inner region are drawn from the same HR2 distribution as that for the sources in the annulus; however, there are 12 such sources.

The diffuse, hard X-ray emission in the Galactic Centre, which dominates the inner 8 pc of the Galaxy⁶, is due to unresolved magnetic cataclysmic variable stars (in particular, a sub-class known as intermediate polars) with optically thin thermal emission of $kT \approx 8\text{--}40 \text{ keV}$ (ref. 13). These intermediate polars are the origin of the $\text{HR2} > 0.3$ population. The newly identified $\text{HR2} < 0.3$ population is prominent only inside the central parsec.

Before concluding that the HR2 probability distribution in the inner parsec is due to a newly identified source population, we studied the confounding effects of the high concentration of gas and dust in the Galactic Centre. The high dust column density could preferentially scatter soft X-rays or lead to the well-known degeneracy between the spectral hardness and interstellar absorption of a source¹⁴. Both of these effects could alter the HR2 distribution near Sgr A*. They can be probed using a second, lower-energy-band hardness ratio to produce colour–colour plots, so-called quantile diagrams, and spectral simulations of thermal and non-thermal sources. All of these analyses demonstrate that the $\text{HR2} < 0.3$ sources cannot be $\text{HR2} > 0.3$ sources masquerading as apparent $\text{HR2} < 0.3$ sources as a result of scattering or column-density effects (see Methods).

The stacked spectra of the $\text{HR2} < 0.3$ and $\text{HR2} > 0.3$ sources in the inner parsec are compared in Fig. 2. The comparison provides further confirmation of the spectrally distinct nature of the $\text{HR2} < 0.3$ sources in the inner parsec. The $\text{HR2} > 0.3$ sources are well fitted by a thermal spectrum with partially covered absorption to account for an accretion curtain and scattering from the surface of a white dwarf—a model that is commonly applied to intermediate polars. The best-fitting temperature ($kT = 6.3^{+1.6}_{-1.7} \text{ keV}$, where the errors here and elsewhere

¹Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, Room 1027, New York, New York 10027, USA. ²Instituto de Astrofísica, Facultad de Física, Pontificia Universidad Católica de Chile, 306, Santiago 22, Chile. ³Millennium Institute of Astrophysics, Vicuña Mackenna, 4860, 7820436 Macul, Santiago, Chile. ⁴Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, Colorado 80301, USA. ⁵Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, MS-83, Cambridge, Massachusetts 02138, USA.

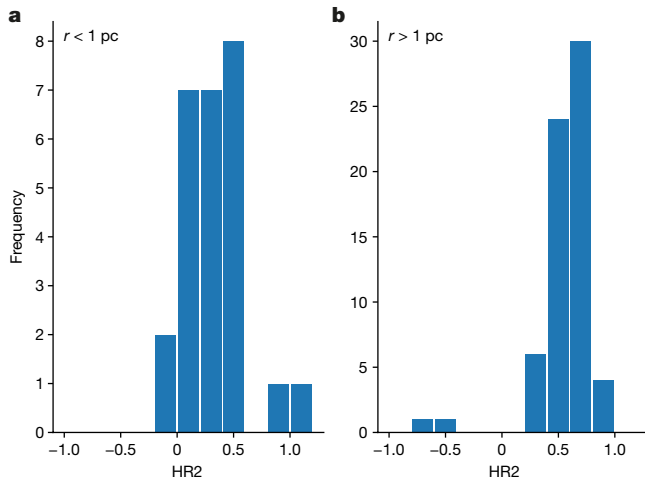


Figure 1 | Hardness ratio (HR2) distribution of X-ray point sources with net counts of $C \geq 100$ at a projected radial distance from Sgr A*. **a**, Sources for $r < 1$ pc. **b**, Sources for $1 \text{ pc} < r < 3.8$ pc. A substantial population of sources with $-0.1 < \text{HR2} < 0.3$ appears for $r < 1$ pc (**a**). Of the 66 sources in the annulus (**b**), 2 (3%) have $\text{HR2} < 0.3$, whereas of the 26 sources in the inner circle (**a**), 12 (46%) have $\text{HR2} < 0.3$. A Kolmogorov–Smirnov test shows that the HR2 cumulative distribution function of the annular region is not consistent with that of the circular region ($D = 0.528$; $P = 3.0 \times 10^{-5}$). The two outliers in **b**, at $\text{HR2} = -0.56$ and $\text{HR2} = -0.74$, were included in the Kolmogorov–Smirnov test. They account for at most one $\text{HR2} < 0.3$ source in **a**, with a probability of about 2%. A χ^2 test based on the lack of spatial constancy in HR2 between $1 \text{ pc} < r < 3.8$ pc and $r < 1$ pc also indicates a different source population for $r < 1$ pc ($P = 1.1 \times 10^{-19}$). The Kolmogorov–Smirnov test, outliers and the χ^2 test (Extended Data Fig. 3) are further discussed in Methods.

correspond to one standard deviation) is consistent not only with the temperatures of the $\text{HR2} > 0.3$ sources at $1 \text{ pc} < r < 3.8$ pc, but also more generally with the low-temperature component of the (unresolved) hard X-ray emission and the typical Chandra-measured temperature found for magnetic cataclysmic variable stars in the Galactic Centre¹⁵. By contrast, the stacked spectrum of the $\text{HR2} < 0.3$ sources in the inner parsec is well fitted (with a reduced chi-squared of $\chi^2_\nu = 0.81$ for 51 degrees of freedom (d.o.f.)) by a non-thermal power-law with $\Gamma = 1.9^{+0.3}_{-0.3}$, with no sign of neutral or ionized iron lines in the spectrum. A final confirmation of the newly identified, spectrally distinct source population in the inner approximately 1 pc comes from

the 13 sources in the central parsec that are bright enough ($C \geq 200$) to enable spectral fitting. The $\text{HR2} < 0.3$ sources are well fitted with a non-thermal spectrum and the $\text{HR2} > 0.3$ sources with an intermediate-polar thermal model (see Methods for discussion and Extended Data Fig. 1 for examples).

The spatially distinct morphology of this newly identified source population is shown in Fig. 3, with the $\text{HR2} < 0.3$ sources concentrated in the central parsec and the $\text{HR2} > 0.3$ sources extended over the region of the diffuse, hard X-ray emission. The newly identified population is not extragalactic because the number of background active galactic nuclei (about 0.1)¹⁶ is negligible within the inner approximately 1 pc above the 2–8-keV flux threshold. The non-thermal spectrum, with $\Gamma \approx 1.5$ –2 and lacking iron lines, clearly distinguishes this population from the population of thermally emitting intermediate polars that dominate at larger radii. We rule out neutron-star, low-mass X-ray binaries as candidates for the newly identified population because of the short recurrence time of about 5–10 years between their large X-ray outbursts. The continuous monitoring of the Galactic Centre for more than a decade is believed to have revealed all such binaries^{17,18}, and none of the point sources discussed here has ever produced an outburst. High-mass X-ray binaries comprise about 40% of all Galactic X-ray binaries¹² and, given the large concentration of O and B stars in the central parsec¹⁹, are potential contributors to the newly identified source population. However, infrared surveys that cover all of the Chandra X-ray sources in the Galactic Centre are able to rule out massive stellar companions^{20,21}, from bright O or B supergiants down to the faintest of the Be high-mass X-ray binaries (spectral class B2V), effectively eliminating the possibility of a contribution of high-mass X-ray binaries. Coronally active isolated stars and stellar binaries having soft thermal spectra (corresponding to $\text{HR2} < -0.3$) are not observed among our sources in the central approximately 1 pc; in addition, their expected luminosities are well below the Chandra detection threshold. During an outburst they may have harder X-ray spectra, but would still be below the Chandra detection threshold, and their numbers are substantial only in the inner approximately 0.1 pc (ref. 22).

The most plausible explanation for the non-thermal sources in the inner parsec is that they are quiescent black-hole low-mass X-ray binaries (qBH-LMXBs). Although qBH-LMXBs can exhibit modest variability²³, there is limited information about their long-term variability, so steady emission from qBH-LMXBs cannot be ruled out. Timing analysis of the non-thermal sources in the inner parsec over the 12 years of observations shows that 6 out of 12 of them are variable, whereas the remaining 6 are steady. Rotation-powered

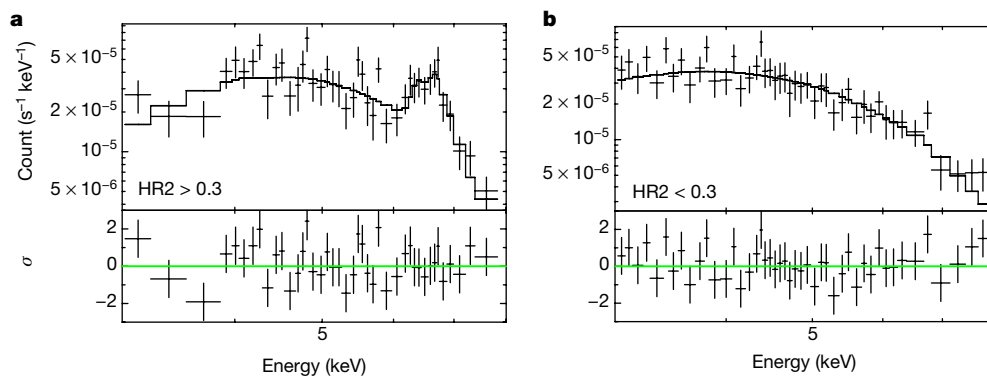


Figure 2 | Stacked Chandra spectra for X-ray sources within the inner parsec. **a**, Sources with $\text{HR2} > 0.3$. **b**, Sources with $\text{HR2} < 0.3$. The $\text{HR2} > 0.3$ spectrum (**a**) for the 8 sources with strong Fe lines is well fitted by an optically thin, thermal plasma model with temperature $kT = 6.3^{+1.6}_{-1.7}$ keV ($\chi^2_\nu = 1.25$ for 36 d.o.f.). A power-law model yields a poor fit ($\chi^2_\nu = 1.90$ for 39 d.o.f.; $\Gamma = 0.7^{+0.5}_{-0.5}$). The 6 sources with weak or no Fe lines have a comparable best-fit photon index of $\Gamma = 0.6^{+0.4}_{-0.4}$. These photon indices are typical of intermediate-polar spectra fitted with a power-law model in the Chandra energy bandpass^{11,13}. The $\text{HR2} < 0.3$

spectrum (**b**) is well fitted by a power-law model with a significantly softer photon index ($\Gamma = 1.9^{+0.3}_{-0.3}$ for 51 d.o.f.). Stacked spectra of the 14 $\text{HR2} > 0.3$ sources (Extended Data Fig. 7) were fitted to a thermal plasma model with the best-fit temperature $kT = 7.3^{+3.0}_{-1.3}$ keV ($\chi^2_\nu = 0.80$ for 58 d.o.f.), the typical Chandra-measured temperature (about 7–9 keV) for magnetic cataclysmic variables in the Galactic Centre⁶. The error bars represent 1σ statistical uncertainties. The bottom panels show residuals (data minus model) in terms of 1σ significance. For more details on the stacked spectra analysis, see Methods.

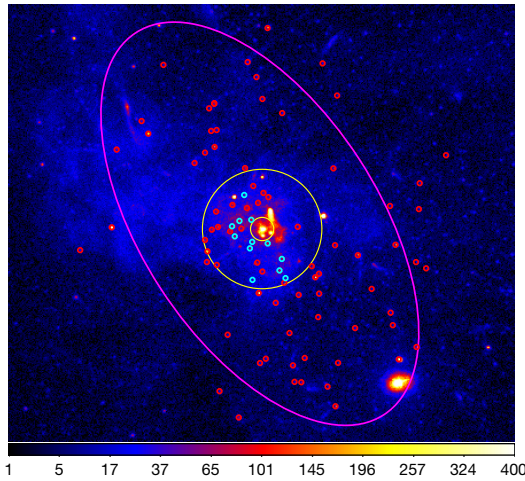


Figure 3 | Chandra 2–8-keV image of the Galactic Centre with X-ray sources with $C \geq 100$ overlaid. Thermal ($\text{HR2} > 0.3$) and non-thermal ($\text{HR2} < 0.3$) sources are indicated with red and cyan circles, respectively. The colour scale indicates the number of 2–8-keV counts per pixel. The inner and outer yellow circles delineate the $0.2 \text{ pc} < r < 1 \text{ pc}$ region around Sgr A*. The inner region was excluded from the analysis. The magenta ellipse ($7.8 \text{ pc} \times 3.9 \text{ pc}$, full-width at half-maximum) bounds the region of spatially unresolved hard X-ray emission that was discovered by NuSTAR and is due to thermal emission from intermediate polars. The non-thermal sources cluster inside the inner parsec, whereas the thermal sources are distributed more uniformly throughout the hard X-ray emission. The Galactic plane runs along the semi-major axis of the ellipse. The dearth of sources to the north of the Galactic plane is due to strong extinction from molecular clouds (such as the circumnuclear disk), which densely populate that region²⁷. This may reduce the estimated number of sources by a factor of a few. Some bright sources without circles are X-ray transients, including the outbursting neutron-star low-mass X-ray binary AX J1745–2901 in the lower right corner. For reference, Chandra’s angular resolution is 0.5 arcsec, which corresponds to about 0.025 pc at the Galactic Centre.

millisecond pulsars (rMSPs) also have non-thermal emission consistent with $\text{HR2} < 0.3$. However, rMSPs are always steady over timescales of months to years²⁴, so we cannot rule out that up to one-half of the non-thermal sources that we observe in the inner parsec are rMSPs.

In Fig. 4 we show the surface density of the 12 non-thermal sources as a function of projected radius from Sgr A*, from which we extract a three-dimensional cusp power-law index. The cusp index does not change in a statistically significant fashion if the six qBH-LMXB candidates are analysed separately from the six rMSP candidates (see Methods for discussion). Owing to the complexities of the formation and evolution of BH-LMXBs, the agreement with the isolated black-hole prediction is probably coincidental. However, the concentration of black holes in the inner approximately 1 pc (well inside the influence radius at $r \approx 3 \text{ pc}$) that we observe has been predicted previously³. In Extended Data Fig. 2 we show the $\log N$ – $\log S$ (number versus X-ray flux) for the qBH-LMXB candidate sources with $r < 1 \text{ pc}$ and down to the Chandra flux limit. A power-law fit to the $\log N$ – $\log S$ distribution ($N(>S) = kS^{-\alpha}$, $\alpha = 1.8^{+0.2}_{-0.2}$) can then be extrapolated to the minimum flux of the local, known BH-LMXB population (see Methods) to give 600–1,000 qBH-LMXBs (1σ error on the fit) if all of the observed $\text{HR2} < 0.3$ sources are qBH-LMXBs, and 300–500 if one-half of them are rMSPs. The lack of large outbursts from individual sources over more than a decade suggests outburst recurrence times of about 1,000 years, for which there is some theoretical motivation (see Methods). If the density cusp extends to Sgr A*, then the qBH-LMXB counts must be corrected for the unobserved volume inside 0.2 pc. For the best-fitting α and three-dimensional cusp index γ , this correction would increase the total number of qBH-LMXBs for $r < 1 \text{ pc}$ by about 50%. To estimate an upper limit for the total number of rMSPs we use a previous result⁶, obtained by using the measured correlation between

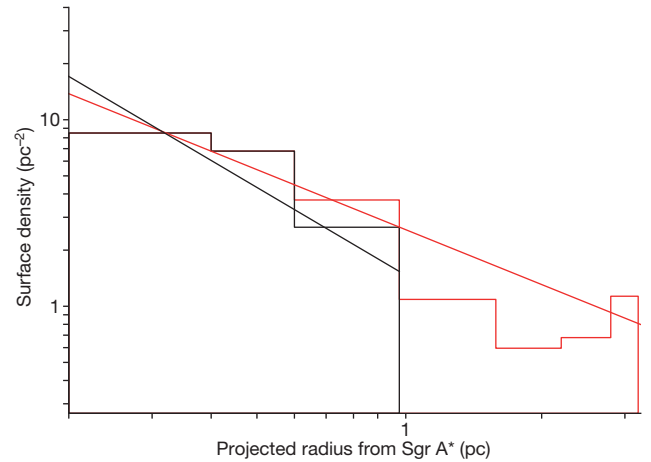


Figure 4 | Surface density of the 12 non-thermal sources as a function of projected radius from Sgr A*. Sources with $C \geq 100$ and $C \geq 50$ are indicated in black and red, respectively. The cusp power-law index γ of the non-thermal sources ($\text{HR2} < 0.3$) was obtained by using an assumed three-dimensional form for the source density of $n(r) = kr^{-\gamma}$ and projecting this along the line-of-sight radius R to obtain a best fit to the surface density $\Sigma(R) = kR^{-\beta}$. The best fit yields $\gamma = 2.4^{+0.3}_{-0.3}$. The isolated black-hole cusp around Sgr A* is predicted to have a power-law index γ in the approximate range 1.3–2.3 (refs 26, 28–30). The red histogram is for sources with $C \geq 50$ and gives a crude estimate of the surface density at larger radii, although it may suffer from mild contamination from spurious background sources. The best-fitting power-law index in this case is $\gamma = 2.0^{+0.2}_{-0.2}$. There is no statistically significant difference in the power-law index if the six rMSP candidates are removed from the analysis. See Methods for more details of the cusp analysis.

the spin-down power \dot{E} and the X-ray luminosity L_X for a large sample of rMSPs, that the fraction of rMSPs above the Chandra flux threshold at the Galactic Centre is about 3%. We therefore set an upper limit of approximately 200 for the number of rMSPs in the Galactic Centre.

Our estimated number of BH-LMXBs is a lower limit for the number of isolated black holes in the Galactic Centre. The total number of low-mass X-ray binaries is consistent with a previous estimate²⁵ based on three-body exchange formation in in-falling globular clusters; however, it has been argued²⁶ that this previous estimate is an overestimate by 10–100. If that is the case, then it is most likely that the black-hole binaries were formed via the tidal capture of old, low-mass stars (either orbiting Sgr A* or formed in the surrounding stellar disk), and the observed number of BH-LMXBs necessitates the existence of more than about 10,000 isolated black holes (A. Generozov, B. Metzger, N. Stone and J. Ostriker, manuscript in preparation).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 September 2016; accepted 6 November 2017.

1. Bahcall, J. N. & Wolf, R. A. Star distribution around a black hole in a globular cluster. *Astrophys. J.* **209**, 214–232 (1976).
2. Bahcall, J. N. & Wolf, R. A. Star distribution around a black hole in a globular cluster. II. *Unequal star masses*. *Astrophys. J.* **216**, 883–907 (1977).
3. Miralda-Escudé, J. & Gould, A. A cluster of black holes at the Galactic Center. *Astrophys. J.* **545**, 847–853 (2000).
4. Morris, M. Massive star formation near the Galactic Center and the fate of the stellar remnants. *Astrophys. J.* **408**, 496–506 (1993).
5. Freitag, M. & Amaro-Seoane, P. Stellar remnants in galactic nuclei: mass segregation. *Astrophys. J.* **649**, 91–117 (2006).
6. Perez, K. *et al.* Extended hard X-ray emission in the inner few parsecs of the Galaxy. *Nature* **520**, 646–649 (2015).
7. Levin, Y. & Beloborodov, A. M. Stellar disk in the galactic center — a remnant of a dense accretion disk? *Astrophys. J.* **590**, L33–L36 (2003).
8. Antonini, F. & Rasio, F. A. Merging black hole binaries in galactic nuclei: implications for Advanced-LIGO detections. *Astrophys. J.* **831**, 187 (2016).

9. O'Leary, R. M., Kocsis, B. & Loeb, A. Gravitational waves from scattering of stellar-mass black holes in galactic nuclei. *Mon. Not. R. Astron. Soc.* **395**, 2127–2146 (2009).
10. Reid, M. J. The distance to the center of the Galaxy. *Annu. Rev. Astron. Astrophys.* **31**, 345–372 (1993).
11. Muno, M. P. *et al.* A catalog of X-ray point sources from two megaseconds of Chandra observations of the Galactic Center. *Astrophys. J. Suppl. Ser.* **181**, 110–128 (2009).
12. Remillard, R. A. & McClintock, J. E. X-ray properties of black-hole binaries. *Annu. Rev. Astron. Astrophys.* **44**, 49–92 (2006).
13. Hailey, C. J. *et al.* Evidence for intermediate polars as the origin of the Galactic Center hard X-ray emission. *Astrophys. J.* **826**, 160 (2016).
14. Hong, J., Schlegel, E. M. & Grindlay, J. E. New spectral classification technique for X-ray sources: quantile analysis. *Astrophys. J.* **614**, 508–517 (2004).
15. Muno, M. P. *et al.* The spectra and variability of X-ray sources in a deep Chandra observation of the Galactic Center. *Astrophys. J.* **613**, 1179–1201 (2004).
16. Kim, M. *et al.* Chandra multiwavelength project X-ray point source number counts and cosmic X-Ray background. *Astrophys. J.* **659**, 29–51 (2007).
17. Degenaar, N. *et al.* The Swift X-ray monitoring campaign of the center of the Milky Way. *J. High Energy Astrophys.* **7**, 137–147 (2015).
18. Degenaar, N. *et al.* A four-year XMM-Newton/Chandra monitoring campaign in the Galactic Centre: analysing the X-ray transients. *Astron. Astrophys.* **545**, A49 (2012).
19. Bartko, H. *et al.* An extremely top-heavy IMF in the Galactic Center stellar disks. *Astrophys. J.* **708**, 834–840 (2010).
20. Mauerhan, J. C. *et al.* Near-infrared counterparts to Chandra X-ray sources toward the Galactic Center. I. Statistics and a catalog of candidates. *Astrophys. J.* **703**, 30–41 (2009).
21. Laycock, S. *et al.* Constraining the nature of the Galactic Center X-ray source population. *Astrophys. J.* **634**, L53–L56 (2005).
22. Sazonov, S., Sunyaev, R. & Revnivtsev, M. Coronal radiation of a cusp of spun-up stars and the X-ray luminosity of Sgr A*. *Mon. Not. R. Astron. Soc.* **420**, 388–404 (2012).
23. Plotkin, R. M. *et al.* The X-ray spectral evolution of galactic black hole X-ray binaries toward quiescence. *Astrophys. J.* **773**, 59 (2013).
24. Bogdanov, S. *et al.* Chandra X-Ray observations of 19 millisecond pulsars in the globular cluster 47 Tucanae. *Astrophys. J.* **646**, 1104–1115 (2006).
25. Muno, M. P. *et al.* An overabundance of transient X-ray binaries within 1 parsec of the Galactic Center. *Astrophys. J.* **622**, L113–L116 (2005).
26. Hopman, C. Binary dynamics near a massive black hole. *Astrophys. J.* **700**, 1933–1951 (2009).
27. Christopher, M. H. *et al.* HCN and HCO⁺ observations of the Galactic circumnuclear disk. *Astrophys. J.* **622**, 346–365 (2005).
28. Alexander, T. & Hopman, C. The effect of mass segregation on gravitational sources near massive black holes. *Astrophys. J.* **645**, L133–L136 (2006).
29. Alexander, T. & Hopman, C. Strong mass segregation around a massive black hole. *Astrophys. J.* **697**, 1861–1869 (2009).
30. Merritt, D. The distribution of stars and stellar remnants at the Galactic Center. *Astrophys. J.* **718**, 739–761 (2010).

Acknowledgements This work was partially supported by NASA contract no. NNG08FD60C. We thank P. Broos of the Chandra ACIS Instrument Team at Penn State for help with the ACIS Extract Package, and D. P. Huenemoerder and H. M. Guenther at MIT for help with the MARX simulation tool. We acknowledge C. Jin and G. Ponti for providing the dust scattering model. We thank A. Generozov, D. Helfand, L. Hui, S. Mandel, B. Metzger, J. Ostriker and N. Stone of Columbia University and Q. D. Wang at the University of Massachusetts at Amherst for discussions. F.E.B. acknowledges support from CONICYT-Chile (Basal-CATA PFB-06/2007, FONDECYT Regular 1141218), the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC120009, awarded to The Millennium Institute of Astrophysics, MAS.

Author Contributions C.J.H., statistical and population analysis, interpretation and manuscript preparation; K.M., image and spectral analysis, interpretation and manuscript preparation; F.E.B., source extraction, MARX simulations and review; M.E.B., image, spectral and statistical analysis, MARX simulations and review; J.H., source extraction, MARX simulations and review; and B.J.H., image, spectral and population analysis, XSPEC simulations and review.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to C.J.H. (chuckh@astro.columbia.edu).

Reviewer Information *Nature* thanks F. Baganoff, J. Miralda-Escudé and M. Morris for their contribution to the peer review of this work.

METHODS

We used data from the Chandra X-ray Observatory (CXO). Following previous survey analyses¹¹ only data from the ACIS-I instrument were used. There were 45 observations pointing at or near Sgr A* providing a total cleaned exposure of 1.4 Ms. This is a 40% increase in exposure time over the most recent and updated (hereafter Muno) source catalogue¹¹, although all of the sources we analysed are listed in the Muno catalogue. 1σ errors are quoted for parameter uncertainties throughout the paper.

Point source selection. We selected all point sources in the Muno catalogue within a circle of radius $r = 100''$ (about 4 pc) centred on Sgr A*. We checked to ensure that these point sources were not embedded in X-ray filaments or molecular clouds. We excluded all known sources of transient outbursts such as neutron-star low-mass X-ray binaries (NS-LMXBs). The region at $r < 5''$ (less than about 0.2 pc) was also excluded because of very high background contamination and owing to the complex X-ray morphology that enhanced difficulties with isolating the true properties of the point sources; this region includes Sgr A*, the IRS 13 star-forming complex and PWN G359.96–0.04. This left 415 point sources from the catalogue out to $r < 100''$. Sources with $C \geq 200$ (41 sources, 10%) after background subtraction could be spectrally fitted individually. Analysis with hardness ratio (colour) was possible for sources with $C \geq 50$ (211 out of 415, 51%). However, most of our analysis was done for sources with $C \geq 100$, because it is increasing difficulty to distinguish between hard and soft sources as the count rate decreases.

Source and background extraction. For each source, ACIS Extract (AE) (version AE 2016Sept22; ref. 31) was used to extract source photons from a circle containing 90% of the encircled energy fraction (EEF), typically with $1''$ radius, centred on each source. All available observations from years 2003–2014 were used. For each Chandra observation, AE provides event and image files for the source, spectral files including source and background spectra, response matrix and effective area files, and light curves. We investigated various methods of background extraction and adopted the following source/background extraction methodology as optimal (and later validated by our MARX simulations) in the crowded Galactic Centre region where background is not spatially uniform. For background subtraction, AE defines an inner annulus at 1.1 times the 99% EEF radius (typically $2.5''$) so that source photons from the wings of the point spread function do not strongly contaminate background estimates. The outer radius of the background annulus is increased until the ratio of its area to the area of the source extraction region reaches an optimum preset value ($= 5$) as recommended in AE. The outer radius of the background annulus can vary over a range of about $3.5''$ to $5''$ because as additional sources are encountered from the inputted region source list, AE automatically excludes those sources and subsequently increases the outer radius until the optimum preset value of background-to-source area ratio is reached. After AE processing, we further excluded point sources with very high background counts (usually from other nearby point sources) from our analysis. Extended Data Fig. 4 illustrates our source and background extractions for one of the 12 soft sources (source 5 in Extended Data Table 1) located in the Rockefeller Ridge, the most source-crowded region we analysed.

Spectral analysis. We applied four different methods of spectral analysis: the Chandra-defined hardness ratio (such as X-ray colour or slope) $HR2 = (C_H - C_L) / (C_H + C_L)$, where C_H and C_L are the net counts in the 2–4-keV and 4–8-keV bands, respectively; colour–colour diagrams using $HR2$ and $HR3 = (C_H - C_L) / (C_H + C_L)$, where C_H and C_L are now the net counts in the 1–3-keV and 5–8-keV bands; spectral fitting of individual sources, with net count $C \geq 200$; and spectral fitting of stacked source spectra. We performed all spectral fitting and calculated hardness ratios using XSPEC version 12.9³². We adopted the *tbabs* model and the previously reported³³ abundance data to take into account interstellar absorption. In addition, the Chandra simulation tool (MARX version 5.3.1) was used extensively³⁴. MARX allows simulated sources with pre-defined spectra to be added directly into the multiple-observation Chandra data that we analysed, with their actual backgrounds, in the Galactic Centre. MARX was used to simulate non-thermal power-law spectra to assess efficiency for recovering the proper photon index and hardness ratios as a function of source brightness, in the actual Chandra Galactic Centre data. Finally, the standard XSPEC simulation tool *fakeit* was used to determine $HR2$ and $HR3$ for various source spectra, to investigate the effects of dust scattering on hardness ratios and power-law photon indices and to investigate any degeneracy between photon index determination and column density.

Hardness ratio analysis. We use hardness ratio analysis to determine whether the point source population is composed of spectrally distinct components. Hardness ratios permit assessment of spectral properties of fainter sources than would be possible with spectral fitting. Previous work¹⁵ has established that $HR2$ is an effective diagnostic for distinguishing magnetic cataclysmic variables (mCVs) with thermal spectra from millisecond pulsars, neutron-star binaries and black-hole binaries with non-thermal spectra. mCVs are known to emit as an optically thin plasma of approximate temperature 8–40 keV with an accompanying iron-line

complex at $E \approx 6$ –7 keV (ref. 13). This is in contrast to the populations of millisecond pulsars, neutron-star binaries and black-hole binaries, which exhibit their true non-thermal power-law spectra with photon indices of $\Gamma \approx 1$ –2.5 (refs 12, 23). On the basis of simulations of Chandra spectra, we adjusted the energy bands of $HR2$ so that the largest distinction between the thermal and non-thermal models could be obtained. Our resultant definition of $HR2$ differs slightly from that used previously¹⁵. $HR2 = 0.3$ was adopted as the dividing line between the thermally emitting mCVs ($HR2 > 0.3$) and the non-thermal source populations ($HR2 < 0.3$). Because higher $HR2$ corresponds to spectrally harder sources, mCVs appear harder in the Chandra bandpass even though they have thermal spectra. More recent work^{6,13,35} has shown that the mCVs in the Galactic Centre are in fact completely dominated by a sub-class of mCV called intermediate polars (IPs); henceforth, we use that designation for the thermal source population characterized by $HR2 > 0.3$. We performed XSPEC simulations to establish the conversion from photon index to $HR2$, for a range of neutral hydrogen column densities typical of the Galactic Centre ($(6$ – $17) \times 10^{22} \text{ cm}^{-2}$; refs 36, 37). XSPEC simulations were also used to determine at what $HR2$ typical IPs with $kT \approx 8$ –40 keV would appear. For this simulation, we chose four canonical IPs (V709 Cas, NY Lup, V1223 Sgr and TV Col), representative of a range of plasma temperatures and neutral hydrogen column densities associated with an accretion curtain at the distance of the Galactic Centre (8 kpc)¹⁰. In all cases the IPs had $HR2 > 0.3$. For reference, $HR2 = 0.3$ corresponds to $\Gamma = 1.5$ for non-thermal spectra with a hydrogen column density of $N_H = 14 \times 10^{22} \text{ cm}^{-2}$, the median of the range of values in the Galactic Centre. This Γ is typical for the majority of NS- and BH-LMXBs, and rMSPs^{12,23,38}.

The $HR2$ distribution for $1 \text{ pc} < r < 3.8 \text{ pc}$ has two extreme outliers (CXO J174540.20–285900.8 and CXO J174537.98–290134.5), at $HR2$ values of $-0.74^{+0.13}_{-0.13}$ and $-0.56^{+0.09}_{-0.09}$, respectively. These sources are not considered to be IPs or non-thermal sources (see below), but they cannot be excluded from the Kolmogorov–Smirnov (KS) test because they lie in a range of $HR2$ where, with their errors, they could contribute to the $r < 1 \text{ pc}$ $HR2$ distribution. However, because these are the only two outliers at $HR2 < 0.3$, out of the 66 sources for $1 \text{ pc} < r < 3.8 \text{ pc}$, there can be at most 1 out of the 12 $HR2 < 0.3$ sources for $r < 1 \text{ pc}$ that arises from the $1 \text{ pc} < r < 3.8 \text{ pc}$ population. This is an extremely conservative estimate. Using the errors on the two $HR2$ outliers and on all of the $r < 1 \text{ pc}$ sources, there is only an approximately 2% chance that even one of the $HR2 < 0.3$ sources for $r < 1 \text{ pc}$ is from the same source population(s) as the $1 \text{ pc} < r < 3.8 \text{ pc}$ sources.

Given the results of the KS test, we performed a parametric test to confirm that the source population for $r < 1 \text{ pc}$ is different from that (or those) for $1 \text{ pc} < r < 3.8 \text{ pc}$. Extended Data Fig. 3 shows the radial distribution of $HR2$ for the sources with $C \geq 100$ as a function of projected distance from Sgr A*. The errors on the $HR2$ determination are normally distributed, as determined by simulations, so a standard χ^2 test can be applied to test for the constancy (expected for an IP population) of $HR2$ as a function of distance from Sgr A*. Fitting a function $HR2(r) = \text{constant}$ to the data for $r > 1 \text{ pc}$ yields the best-fitting $HR2$ of 0.59. The fact that the reduced χ^2 value ($\chi^2_\nu = 1.39$ for 63 d.o.f.) exceeds 1 suggests that the spectral hardness data are intrinsically scattered, probably owing to the spread of the IP temperature distribution, which is typically in the range $kT \approx 8$ –40 keV. We estimated the systematic errors associated with the intrinsic $HR2$ distribution of $r > 1 \text{ pc}$ sources by taking the Gaussian quadrature difference between the total variances and the statistical errors. We found that the systematic errors (about 8%) are subdominant compared to the statistical errors (about 20%). To test the hypothesis of a bifurcation in the $HR2$ distribution at $r < 1 \text{ pc}$, we fitted a constant $HR2$ function to the sources for $0.2 \text{ pc} < r < 3.8 \text{ pc}$, taking into account the systematic and statistical errors. The fit gives a large χ^2 value ($\chi^2_\nu = 3.0$ for 89 d.o.f., $P = 1.1 \times 10^{-19}$), which establishes the non-constancy of $HR2$ and thus the presence of a different source population for $r < 1 \text{ pc}$. The outliers were excluded in this analysis of $HR2$ spatial constancy because their $HR2$ values are more than 5σ from even the lowest $HR2$ (with error) from the 64 other $1 \text{ pc} < r < 3.8 \text{ pc}$ sources. However, consistent with the KS test, one $HR2 < 0.3$ source at $r < 1 \text{ pc}$ could be associated with the $r > 1 \text{ pc}$ source population(s). Because the χ^2 test is parametric and depends on the assumption of spatial constancy of the dominant IP population, it is weaker than the non-parametric KS test, which makes no assumptions about the spatial dependence of $HR2$ and includes the outliers; but it provides additional confirmation of the newly identified population for $r < 1 \text{ pc}$.

The two outlier sources have sufficient counts to enable us to fit their Chandra spectra. Fitting with an absorbed power-law model yields very soft photon indices ($\Gamma > 4$), indicating thermal X-ray spectra. The source CXO J174540.20–285900.8 fits well to an absorbed blackbody model with $kT \approx 0.5 \pm 0.1 \text{ keV}$. The best-fitting blackbody temperature is consistent with persistent thermal emission of a magnetar with typical temperatures of $kT \approx 0.3$ –0.5 keV (ref. 39). The other source (CXO J174537.98–290134.5) exhibits several emission lines in the energy range $E \approx 2$ –3 keV and its spectrum fits to an absorbed, optically thin thermal plasma (*tbabs***apec*) model with best-fitting temperature $kT = 0.8 \pm 0.1 \text{ keV}$.

This source is probably a dwarf nova, given that its temperature is lower than those of IPs and polars. The two outliers are not shown in Fig. 3 and Extended Data Fig. 3. **Colour–colour analysis.** One systematic uncertainty that must be addressed in both the hardness ratio analysis and the spectral analysis is the possible degeneracy between hydrogen column density and spectral hardness/softness. There are spatial variations in the hydrogen column density in the Galactic Centre. These variations can cause an intrinsically hard source at low N_{H} in the Chandra bandpass to be misidentified as a more highly absorbed soft source, because both may yield the same HR2. We investigated whether this degeneracy could lead to systematically lower HR2 values in the central parsec. The degeneracy would also affect the spectral analysis, as discussed below. To evaluate the possibility of degeneracy affecting our analysis, we resorted to colour–colour plots, a method that is commonly used to estimate the severity of absorption effects. For this investigation the second colour HR3 was used, because it utilizes the 1–3-keV energy band and so is more sensitive to hydrogen column density than is HR2, which focuses on higher-energy X-rays. To assess the sensitivity of HR3 to absorption effects, the XSPEC routine *fakelr* was used to generate simulated Chandra spectra of both IP thermal spectra at $kT = 8\text{--}40$ keV and non-thermal power-law spectra with $\Gamma \approx 1\text{--}2.5$. XSPEC simulations were used to calculate HR2 and HR3 for the input spectra. They demonstrated that an intrinsic HR2 > 0.3 source could only appear as an apparent HR2 < 0.3 source, owing to lower column density, if HR3 ≤ 0.6 . On the contrary, HR2 < 0.3 sources always have HR3 ≥ 0.9 . Extended Data Fig. 5 shows a HR2–HR3 scatter plot of all of the sources, differentiating those for $r < 1$ pc (cyan) and those for $r > 1$ pc (red). With the exception of two sources (which are intrinsically hard sources with low $N_{\text{H}} \approx 10^{22} \text{ cm}^{-2}$, on the basis of spectral fitting), all of the HR2 < 0.3 sources have high HR3, confirming that they are intrinsically soft spectral sources rather than hard sources (intrinsic HR2 > 0.3) with low column density masquerading as low-HR2 soft sources. In fact, the systematics of the absorption are such that very high column densities would cause intrinsically low-HR2 sources to appear as higher-HR2 sources with high HR3. Thus the HR2 analysis of the number of soft sources for $r < 1$ pc is conservative.

Quantile hardness ratio analysis. More recently, extensive use has been made of quantile analysis^{14,35} instead of the colour–colour analysis of Extended Data Fig. 5. Quantile plots have been shown to be more effective in breaking photon index (hardness ratio)–hydrogen column density degeneracies. Extended Data Fig. 6 shows the same data as in Extended Data Fig. 5, but in a quantile plot. Here Q_n is the energy within which $n\%$ of the photons in the 2–8-keV band are contained and m is the median energy below which 50% of the photons in the band are contained. The quantile diagram also shows contours of column density and photon index, along which the quantile parameters of the x and y axis have been calculated using an assumed non-thermal spectral shape and Chandra’s response function. The diagram clearly shows the two distinct populations of sources. Nearly all of the sources for $r > 1$ pc have $\Gamma < 1$ whereas about half of the sources for $r < 1$ pc have softer power-law photon indices of $\Gamma \approx 1\text{--}2$. The large scatter of all of the sources with column density shows no systematic bias in source hardness due to column density effects.

Spectral analysis of individual, bright Chandra sources. A spectral analysis, albeit available for only a smaller number of sources (13 for $r < 1$ pc and 28 for $1 \text{ pc} < r < 3.8$ pc) provides a second method for confirming the presence of distinct source populations inside and outside 1 pc. We fitted Chandra spectra of the sources with sufficient net counts ($C \geq 200$) using thermal (*tbabs*pcfabs*apec*) and non-thermal models (*tbabs*powerlaw*) in XSPEC. Spectral bins were adaptively re-binned so that the counts in each bin have at least 3σ statistical significance over background.

Outside $r = 1$ pc, 16 of the 28 HR2 > 0.3 sources with $C \geq 200$ are well fitted by a thermal model with $kT \approx 5\text{--}10$ keV. These brighter sources, as demonstrated by the example shown in Extended Data Fig. 1a, clearly exhibit broad iron lines at 6–7 keV, characteristic of mCVs. The fainter 10 HR2 > 0.3 sources are well fitted by hard power-law models with $\Gamma < 1$. Previous studies¹⁵ with Chandra also found that faint mCVs are readily fitted by power laws with $\Gamma \approx 0\text{--}1$.

In contrast, inside $r = 1$ pc, the five HR2 < 0.3 sources with $C \geq 200$ are well fitted by a power-law model with $\Gamma \approx 1.5\text{--}2$ and no apparent iron lines. Extended Data Fig. 1b shows the spectrum of a bright HR2 < 0.3 source with $C = 267$ fitted to a power-law model with $\Gamma = 1.9 \pm 0.3$. Of the seven HR2 > 0.3 sources, similarly to the case of $r > 1$ pc sources, the five bright sources are well fitted by a thermal model with broad iron lines and the two fainter sources by a power-law model with $\Gamma < 1$.

Using the $\log N\text{--}\log S$ values measured for both of the soft sources, and a much larger sample of hard sources (IPs) in the Galactic Centre ($N(>S) = kS^{-\alpha}$, $\alpha = 1.4$)³⁵, we expect around four soft and eight hard sources inside $r = 1$ pc to have sufficient counts for spectral analysis. These numbers are roughly consistent with the five soft and seven hard sources inside $r = 1$ pc that we found that are amenable to spectral fitting.

MARX simulation. To validate our source/background extraction method and the subsequent spectroscopy and colour photometry, we employed MARX simulations followed by AE analysis. MARX simulation allows us to input sources, representing mCVs and BH-LMXBs, on top of the diffuse, gaseous structures in order to assess whether we can recover the proper input spectrum and the input HR2 hardness ratios within statistical uncertainties. We generated fake point sources with input power-law spectra with $\Gamma = 1.5\text{--}2$. Input X-ray fluxes were adjusted to give $C = 100\text{--}200$ in the 2–8-keV band, similar to those of the 12 HR2 < 0.3 sources for $r < 1$ pc (Extended Data Table 1). In each MARX simulation, 5–10 fake sources were added to all 45 Chandra observations at various source-free locations within $0.2 \text{ pc} < r < 1 \text{ pc}$ including the ridge detected by ref. 40 (representing one of the highest background regions in $r = 0.2\text{--}3.8$ pc). This is the most realistic simulation because we take into account the actual background, vignetting, point spread function and dithering effects for all 45 Chandra observations used for data analysis. A total of 124 hard and soft sources were run at 32 positions at $r < 1$ pc. The MARX sources were placed as close as possible to the observed sources whenever feasible. Often a MARX simulation was run repeatedly at a given position to build up statistics on photometry and spectral recovery in the face of counting statistics. We analysed the simulated Chandra data using AE, calculated hardness ratios of the fake sources and performed spectral fitting with XSPEC. We recovered the (soft) input HR2 values within statistical errors for 65 out of 67 simulated sources with $C \gtrsim 100$. Thus, the probability of misidentifying an intrinsically soft source (HR2 < 0.3) as a hard source (HR2 > 0.3) is about 3%. More importantly, the probability for hard sources (with IP thermal spectra) with $C \gtrsim 100$ being misidentified as soft sources was about 2% (56 out of 57 simulated hard sources). Thus, at $r < 1$ pc, the mean number of soft sources turning into hard sources is about 0.4 and the mean number of hard sources turning into soft sources is about 0.3. For fainter sources with $C \leq 50$, the false probability increases to more than about 20%. Thus, we conclude that the threshold for reliable, spectroscopic identification between non-thermal X-ray sources and IPs is $C \gtrsim 100$ throughout the $0.2 \text{ pc} < r < 3.8$ pc region we analysed.

Scattering effects on spectral analysis. Owing to dust scattering of X-ray photons, a point source can appear with a diffuse halo that extends beyond a typical source-extraction region⁴¹. Dust scattering haloes have been detected from Chandra observations of bright X-ray transients (such as AX J1745.6–2901⁴² and Swift J174540.7–290015⁴³). Because the dust scattering is more substantial at low energy, the X-ray spectrum of a source may appear harder if scattered photons are not fully collected within an extraction region. For example, a non-thermal power-law X-ray spectrum could be hardened by up to $\Delta\Gamma \leq 0.5$ when $N_{\text{H}} > 10^{23} \text{ cm}^{-2}$ (ref. 44). Unlike bright X-ray transients, no dust scattering halo is observed from the (fainter) Chandra X-ray sources because it is buried under diffuse X-ray background in the Galactic Centre.

To assess the effects of dust scattering, we applied a spectral model⁴². The model parameters, such as grain sizes and types, column densities and distances of dust layers, have been uniquely determined by fitting the Chandra radial profiles of the outbursting NS-LMXB AX J1745.6–2901, which lies 1.5 arcmin away from Sgr A*. The model requires a foreground dust layer in the spiral arms a few kiloparsecs away from the Galactic Centre and another layer closer to AX J1745.6–2901. The foreground dust layer with $N_{\text{H}} \approx 1.7 \times 10^{23} \text{ cm}^{-2}$ is likely to lie in the line of sight for other X-ray sources within a few arcmin of Sgr A* (ref. 42).

We simulated Chandra ACIS-I spectra that correspond to our source-extraction size (typically $1''$) using various input spectral models, including a power-law with $\Gamma \approx 1\text{--}2.5$ (representing non-thermal sources) and a partially covered APEC model with $kT \approx 20$ keV (representing typical IP-like spectra). For each spectral model, we simulated Chandra ACIS-I spectra with and without the multiplicative *fgcdust* scattering model in XSPEC. Then, we calculated HR2 hardness ratios. We also fit power-law (*tbabs*powerlaw*) and thermal (*tbabs*pcfabs*apec*) models to the simulated spectra of non-thermal and thermal sources, respectively. The differences in the best-fitting HR2 values, power-law indices and temperatures between the simulated Chandra spectra with and without the dust scattering model provide an estimate for how much spectral hardening could be caused by typical dust scattering in the Galactic Centre. Regardless of the input model, thermal or non-thermal, dust scattering hardens HR2 by $\Delta\text{HR2} = 0.1$. Thus, our hardness selection using HR2 is relatively unaffected by dust scattering for the relevant spectral types. The net effect of dust scattering is to harden non-thermal and thermal Chandra spectra by $\Delta\Gamma = 0.3$ (consistent with previous results⁴⁴) and $\Delta kT = 0.5$ keV, respectively.

Stacked spectra analysis. That the HR2 cut produces two distinct source populations is further supported by an analysis of stacked Chandra ACIS spectra in the central parsec. Using the *combine_spectra* command in the Chandra analysis software CIAO, we combined source and background spectra files for each group of the HR2 < 0.3 and HR2 > 0.3 sources. We excluded several sources from the

stacked spectra because their very high background counts are dominant and so lower the signal-to-noise ratios. Figs 2a and b in the main text show the stacked spectrum of the 8 HR2 > 0.3 sources with strong Fe lines and 12 HR2 < 0.3 sources, respectively. In both cases, all sources have net counts $C \gtrsim 100$ and they are located in the central parsec.

First, we fitted an absorbed power-law model (*tbabs*powerlaw*) to both stacked Chandra spectra. To accurately determine the spectral parameters and column density N_{H} , we applied the multiplicative *fgcdust* dust scattering model⁴¹. The stacked Chandra spectra of the HR2 < 0.3 sources in the central parsec are well fitted ($\chi^2_{\nu} = 0.81$, d.o.f. = 51) by an absorbed power-law model yielding $\Gamma = 1.9^{+0.3}_{-0.3}$ and $N_{\text{H}} = 0.8^{+0.2}_{-0.2} \times 10^{23} \text{ cm}^{-2}$ (Fig. 2b).

We found that the 14 HR2 > 0.3 sources at $r < 1 \text{ pc}$ divide into two subgroups with different spectral characteristics: 8 of them exhibit strong Fe lines in their Chandra spectra, while the other 6 sources show either no or weak Fe lines. An absorbed power-law model fit to both stacked spectra yields a hard photon index of $\Gamma \approx 0.6\text{--}0.7$. The hard photon index ($\Gamma \lesssim 1$) is a common signature of IPs⁴⁵. The power-law model yields a poor fit to the stacked spectra of the 8 HR2 > 0.3 sources with strong Fe lines ($\chi^2_{\nu} = 1.90$ for 39 d.o.f.).

Since the IP spectra are physically best modelled as emission from an optically thin thermal plasma, we employed the APEC model in XSPEC. We added a partial-covering absorption model (*pcfabs*) to account for photo-absorption in the accretion curtain or X-ray reflection from the white dwarf surface and a Gaussian line component for the neutral Fe-K fluorescence line¹³. Using this fiducial IP spectral model, that is, *fgcdust*tbabs*pcfabs*(apec+gauss)*, the stacked Chandra spectra of the 8 HR2 > 0.3 sources with strong Fe lines fit well, yielding the best-fit temperature of $kT = 6.3^{+1.6}_{-1.7} \text{ keV}$ (Fig. 2a), while the stacked Chandra spectra of the 6 HR2 > 0.3 sources with no Fe lines fit to a significantly higher temperature ($kT = 19_{-10} \text{ keV}$). In the latter case, the temperature is not well constrained owing to the lack of Fe lines in the Chandra energy band below 8 keV. It is likely that the HR2 > 0.3 sources are composed of 8 IPs with $kT \approx 5\text{--}8 \text{ keV}$ (with Fe lines) and 6 IPs with higher temperatures where the Fe complex is over-ionized.

The overall stacked spectrum of all 14 HR2 > 0.3 sources fits to a partially covered thermal model well ($\chi^2_{\nu} = 0.80$, d.o.f. = 58), yielding the best-fit temperature of $kT = 7.3^{+3.0}_{-1.3} \text{ keV}$ (Extended Data Fig. 7). The best-fit Fe abundance $A_{\text{Fe}} = 0.6^{+0.3}_{-0.2}$ (relative to solar) is consistent with previous spectral fits for the unresolved, diffuse IP component of the CHX^{6,13}. Our results from the best-fit models are summarized in Extended Data Table 2. The best-fit plasma temperature of $kT = 7.3^{+3.0}_{-1.3} \text{ keV}$ represents an ensemble of IPs with different temperatures and it is consistent with the stacked spectra of the HR2 > 0.3 sources for $1 \text{ pc} < r < 3.8 \text{ pc}$ ($kT = 5.5^{+0.6}_{-0.3} \text{ keV}$) as well as the results from ref. 6 ($kT = 7.5^{+1.5}_{-1.3} \text{ keV}$) and ref. 15 ($kT \approx 7\text{--}9 \text{ keV}$).

Hard and soft spectrum background systematics. Here we address whether the spectral difference between the soft and hard sources could be due to non-uniform background in the Galactic Centre region or to thermal hotspots in the diffuse emission. Extended Data Fig. 8 shows the stacked background spectra for both the HR2 < 0.3 sources and HR2 > 0.3 sources (with $C \geq 100$). They do not fit a power-law model with large $\chi^2_{\nu} \approx 3.4\text{--}3.6$ owing to the emission lines at 2–4 keV (from low-Z elements) and 6–7 keV (from Fe). Both background spectra fit well with an absorbed two-temperature APEC model yielding $\chi^2_{\nu} \approx 1.2\text{--}1.4$. The best-fit temperatures are consistent between the two background spectra: $kT_1 = 0.93^{+0.10}_{-0.04} \text{ keV}$ and $kT_2 = 5.9^{+1.5}_{-2.9} \text{ keV}$ (for the soft source background) and $kT_1 = 1.18^{+0.05}_{-0.04} \text{ keV}$ and $kT_2 = 6.6^{+1.5}_{-1.1} \text{ keV}$ (for the hard source background). Thus the spectrally distinct nature of the HR2 < 0.3 and HR2 > 0.3 sources cannot be attributed to spectral variations in the background, and their backgrounds are not consistent with non-thermal spectra.

Spurious power-law spectra could result from subtracting thermal background spectra from ‘hotspots’¹¹. These point-like hotspots are possibly due to small-scale density variations (about 0.1 pc) in the (thermal) diffuse emission in the Galactic Centre. They are speculated¹¹ to be the origin of a thermal point source excess for low-flux sources in the Galactic Centre. These should become significant only at fluxes that corresponds to $C \leq 40$, about 3–7 times below the net counts of the HR2 < 0.3 sources. Nevertheless, we investigated the possibility that subtracting a similar background spectrum from a hotspot thermal spectrum could produce a power-law spectrum. Using the measured background for the HR2 < 0.3 sources, XSPEC simulations were run for simulated thermal sources with plasma temperatures ($kT_2 = 3.5\text{--}8 \text{ keV}$) and abundance ($Z_2 = 0.3\text{--}1.5$) encompassing those observed in the central few parsecs. The source normalization was adjusted to correspond to our observed source net counts. Using the same rebinning and fitting procedures as for our overall analysis, we found that the stacked spectrum for the 12 simulated thermal hotspots was indeed a power law, but with best-fit photon indices of $\Gamma = 2.5\text{--}4$, steeper than our observed HR2 < 0.3 sources ($\Gamma = 1.9^{+0.3}_{-0.3}$). We could only get closer to our observations by adjusting the net counts to be well

below our 100-count threshold. And in almost all cases including the best-fit photon indices $\Gamma \approx 2.5$, the spectra showed strong emission lines at 2–4 keV and 6–7 keV (see typical example in Extended Data Fig. 9). Consequently, all the fits were poor with an absorbed power-law model ($\chi^2_{\nu} \approx 1.4\text{--}1.8$ with approximately 50–60 d.o.f.) and, owing to the emission lines, looked nothing like our featureless spectra of the HR2 < 0.3 sources.

Variability study. For the 12 HR2 < 0.3 point sources within $r < 1 \text{ pc}$, we evaluated their X-ray variability by applying the Bayesian-Block (BB) algorithm⁴⁶ to the unbinned photon arrival times after removing time gaps between the Chandra observations spanning around 12 years. No exposure map or vignetting effect correction is required because we verified that none of the sources at $r < 1 \text{ pc}$ are near the chip gaps and all 45 observations were pointed at Sgr A*. The BB algorithm is more sensitive to short-term variability than the KS test, which evaluates the statistical significance of variability over the full time interval. Our analysis followed ref. 46. Among the 12 HR2 < 0.3 X-ray sources, 6 sources (50%) were variable at the 90%–99% confidence level (Extended Data Table 1). Since the BB algorithm works on source and background events, we did an independent timing analysis on nearby regions to ensure that variability was not due to time varying background.

Although BH-LMXBs and rMSPs both exhibit non-thermal X-ray spectra, X-ray variability can be used to distinguish between them. Most (around 70%) of the known BH-LMXBs observed in the quiescent state on multiple occasions have shown evidence of X-ray flux variability, many by a factor of 2–5 over a timescale of days to years, owing to accretion²³. On the other hand, rMSPs display no such long-term variability. Whereas a majority of rMSPs emit thermal X-rays with blackbody temperatures of $kT < 0.3 \text{ keV}$ originating from the polar caps, non-thermal X-ray emission has been detected from about 30% of rMSPs^{24,47,48}. The population of rMSPs with only soft thermal emission, if they are located in the Galactic Centre, are not observable in the X-ray band because the soft thermal X-ray emission of this population will be completely absorbed owing to high column density $N_{\text{H}} \approx 10^{23} \text{ cm}^{-2}$. The population with non-thermal emission, which represents just over one-third of all approximately 50 known rMSPs detected in the X-ray band, exhibit power-law spectra with $\Gamma = 1\text{--}1.5$ and $L_{\text{X}} \approx 10^{30}\text{--}10^{33} \text{ erg s}^{-1}$. The non-thermal X-ray emission from rMSPs originates either from the magnetosphere or from an intra-binary shock (for rMSPs in binary systems)^{24,48}. Orbital variability has sometimes been detected owing to eclipsing from the companion⁴⁸. Isolated MSPs do not show X-ray variability on timescales greater than their millisecond spin periods because non-thermal X-rays are emitted from the magnetosphere²⁴. On the other hand, non-thermal X-ray emission from binary rMSPs often exhibits modulation over the orbital periods of these sources (hours to a few days); but these sources do not show the long-term variability that is observed for many of the Chandra point sources in the Galactic Centre. Thus, the detection of long-term variability over months or years in the Chandra sources can be used as a strong indicator against rMSPs. On the other hand, timing studies of nearby qBH-LMXBs show that about 30% of them do not show variability and one would expect an even larger fraction to appear ‘steady’ at the great distance of the Galactic Centre, where variability is harder to detect and poorly constrained in fields of high background. Based on this variability argument, one-half or less of the HR2 < 0.3 sources may be rMSPs.

Population analysis. Flaring emission from the coronae of main-sequence stars has been proposed previously to account for a substantial fraction of the soft-X-ray emission in the Galactic Centre, but only in the central approximately 0.1 pc (ref. 22). The mean luminosity of these stars in the quiescent state is more than three orders of magnitude below our sensitivity limit and, even when flaring, a factor of about 30 below our sensitivity limit ($10^{28}\text{--}10^{30} \text{ erg s}^{-1}$)²². The active stellar binaries (RS CVn, BY Dra and Algols) are similarly excluded because even objects in the bright RS CVn class have a luminosity of about $10^{30}\text{--}10^{31} \text{ erg s}^{-1}$, more than about three times below the detection threshold²². Moreover, because the stellar mass inside 1 pc is 40% less than that in the $1 \text{ pc} < r < 3.8 \text{ pc}$ annulus and X-ray emitting stars are not present in the annulus, the contribution of stars inside 1 pc must be negligible.

Spatial morphology. We extracted the power-law index for the three-dimensional density $n(r)$ of soft (HR2 < 0.3) sources by using the assumed form $n(r) = kr^{-\gamma}$ and projecting this along the line of sight to obtain a best fit to the surface density $\Sigma(R) = kR^{-\beta}$. The appropriate integral is $\Sigma(R) = 2 \int r \rho(r) (r^2 - R^2)^{-1/2} dr$, where r and R are the actual and projected distances from Sgr A*, respectively. We used a Python fitting routine in the *scipy.optimize.curve_fit* package. For the qBH-LMXB candidates this gives $\gamma = 2.4^{+0.3}_{-0.3}$. If the steady rMSP candidate sources are removed, a slightly shallower cusp results, with $\gamma = 2.0^{+0.7}_{-0.7}$. We also included fainter sources with $C \geq 50$ to estimate the radial extent of the soft source distribution without the rMSP candidates. This yields a power-law index of $\gamma = 1.5^{+0.5}_{-0.5}$. The stellar cusp was recently detected⁴⁹ in the population of old stars in the Galactic Centre and at a comparable projected distance range to that explored here, but had an associated power-law fit with a comparable cusp of $\gamma = 1.5$.

Total number of non-thermal sources in the central parsec. We used the log N –log S plot in Extended Data Fig. 2 to estimate the total number of non-thermal sources. The extrapolation to fainter sources below the Chandra flux threshold was based on a previously reported intrinsic luminosity distribution⁵⁰. The slope of this distribution ($\alpha = 1.4^{+0.1}_{-0.1}$) is in agreement, within errors, with that measured for the qBH-LMXB candidates in the central parsec ($\alpha = 1.8^{+0.2}_{-0.2}$). A cross-check on the log N –log S results can be obtained by using the measured slope for the soft, $r < 1$ pc sources to calculate the mean luminosity per source, assuming that the log N –log S plot extends down to the faintest qBH-LMXB observed⁵⁰ locally. We have $L(\text{total})/N = \int N(L)LdL / \int N(L)dL$, where $N(L) = dN(>L)/dL$ and the limits of integration are set by the previously used value⁴⁸ of L_{min} and the intrinsic luminosity of the brightest soft source that we observed in the central parsec. This gives $\langle L \rangle = 4 \times 10^{31} \text{ erg s}^{-1}$ for the soft sources of the central parsec, in good agreement with the previously observed⁵⁰ mean luminosity of $5 \times 10^{31} \text{ erg s}^{-1}$. However, the estimate of the total non-thermal source population should be viewed with caution. It is based on the value of L_{min} of a small sample (18) of relatively nearby black-hole binaries and it is unclear how representative these are of the qBH-LMXBs of the Galactic Centre. As an example, increasing L_{min} by a factor of two would reduce the extrapolated number of sources by a factor of about three.

Although we cannot rule out that the steady sources that comprise one-half of our non-thermal sources in the central parsec are rMSPs, we might expect them to have a more extended spatial distribution, because neutron stars receive substantial kick velocities on formation. Even when situated in binaries, and taking into account the higher escape velocity in the Galactic Centre, a more extended distribution, beyond 1 pc, for the rMSP population is likely. However, because rMSPs are an old population, those that have escaped the centre may be located at much larger distances than the 3.8 pc to which our analysis extends. Those in the central parsec are the fraction that did not have sufficient kinetic energy to escape. In fact, all of our observations are within the approximately 3-pc influence radius in the Galactic Centre. Detailed modelling is required to resolve these questions.

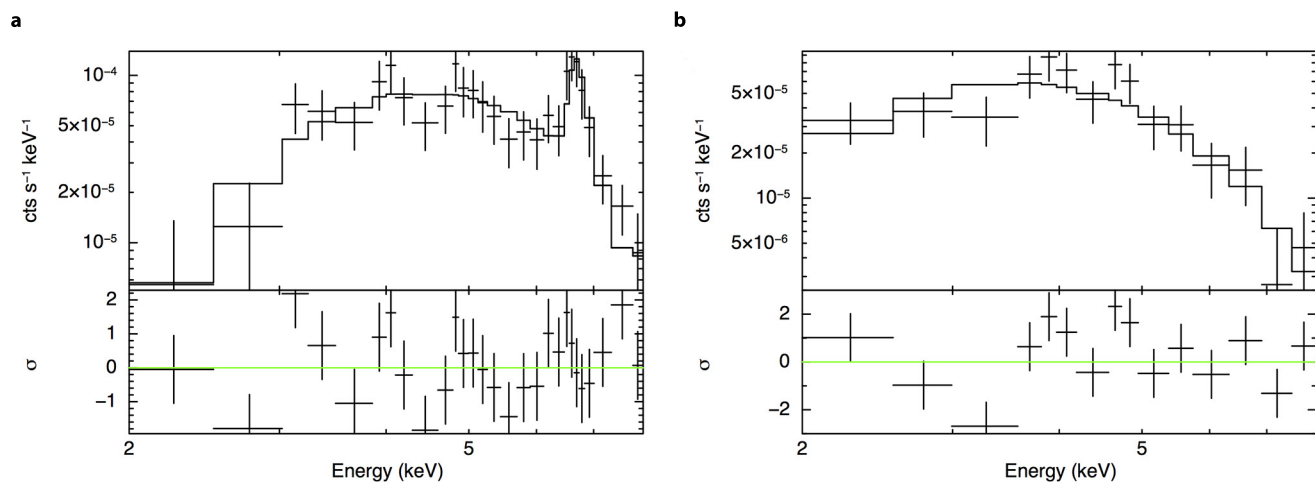
Transient emission from the qBH-LMXB. Four X-ray binary transients were previously reported in the central parsec, representing an over-density per unit stellar mass of about 20 times compared to the surrounding tens of parsecs²⁵. Subsequent work established that three of those transients are NS-LMXBs, leaving one potential BH-LMXB candidate in the central parsec, which has had only its original large outburst in more than a decade of monitoring^{17,18}. More recently, several new transients that are strong BH-LMXB candidates have been identified in the Galactic Centre (K.M., C.J.H., D. Haggard, B.J.H., C. Jin, S. Mandel, M. Nynka, G. Ponti and J. Tomsick, manuscript in preparation). If qBH-LMXBs have large outburst recurrence times of about 100 years, then in more than 10 years of monitoring 40 or more outbursts would have been expected from the extrapolated qBH-LMXB population, which is clearly inconsistent with observations. If the recurrence time is about 1,000 years, then there is rough consistency between the number of BH-LMXB transients observed in the central parsec over the more than 10 years of monitoring, and the non-detection of large outbursts among the dozen X-ray binaries reported here. Such a long recurrence time may not be implausible. Observations of the local qBH-LMXB population reveal a correlation between luminosity and orbital period, such that the sources below our flux threshold all have orbital periods P of less than about 10 h. But according to the accretion disk instability model, very faint systems with P less than about 10 h are precisely those whose accretion rate is below the critical rate required to induce large outbursts. If they are accreting at around 10^{-3} times the critical accretion rate, then recurrence times of around 1,000 years are realized. The qBH-LMXBs above our detection threshold are both brighter and, according to the L_X – P correlation, should have longer periods. They would therefore burst at a higher rate, but still

be consistent with the non-observation of large outbursts from their locations. A similar argument, based on a low accretion rate leading to a lack of recurrent outbursts, has been used to explain the inefficiency of X-ray searches in uncovering black hole–Be binaries⁵¹. No such problem arises for rMSPs, which do not have outbursts.

Code availability. We used standard software for all our data analysis: the Chandra Interactive Analysis of Observations (CIAO; <http://cxc.harvard.edu/ciao/>), the MARX simulator (<http://space.mit.edu/CXC/MARX/>), the ACIS Extract package (<http://personal.psu.edu/psb6/TARA/AE.html>) and NASA's HEASOFT software (<https://heasarc.nasa.gov/lheasoft/>).

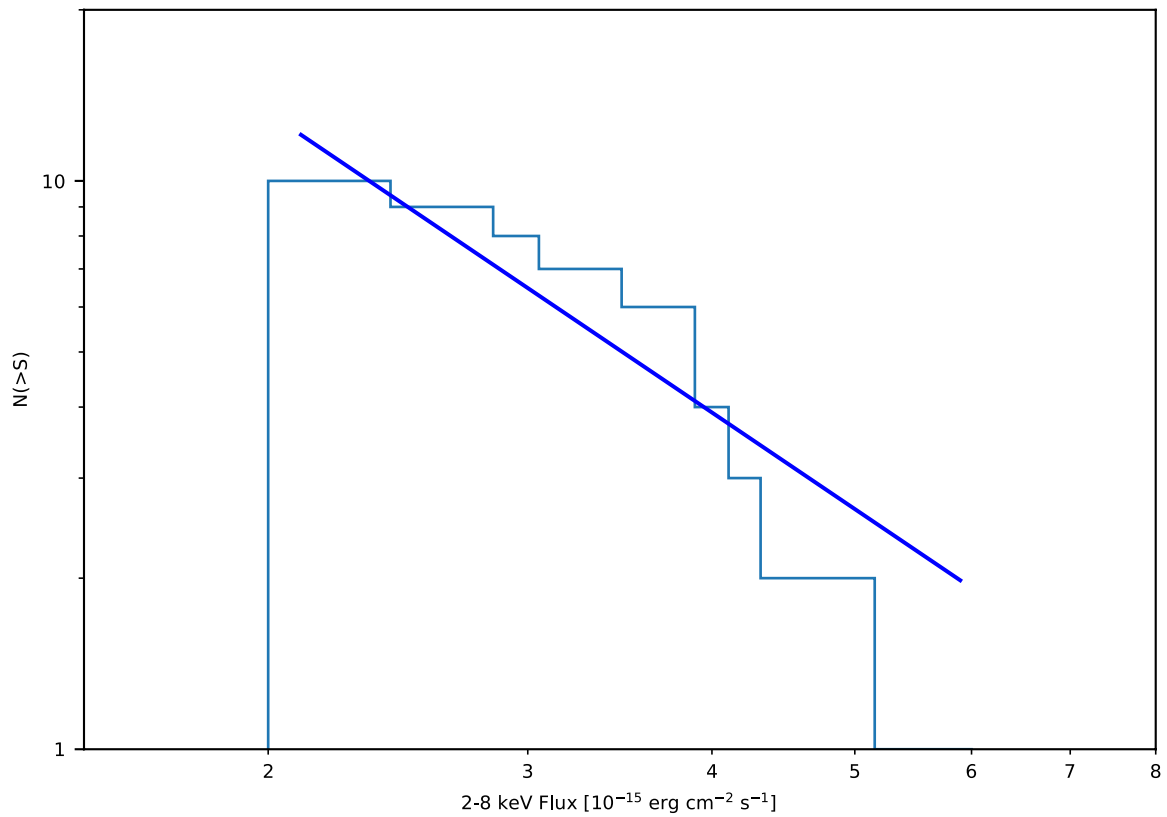
Data availability. The Chandra data that support the findings of this study are available from NASA's HEASARC data archive (<https://heasarc.gsfc.nasa.gov/docs/archive.html>). The data shown in the figures and tables are available from the corresponding author on reasonable request.

31. Broos, P. S. *et al.* Innovations in the analysis of Chandra-ACIS observations. *Astrophys. J.* **714**, 1582–1605 (2010).
32. Arnaud, K. A. XSPEC: the first ten years. *ASP Conf. Ser.* **101**, 17–20 (1996).
33. Wilms, J., Allen, A. & McCray, R. On the absorption of X-rays in the interstellar medium. *Astrophys. J.* **542**, 914–924 (2000).
34. Davis, J. E. *et al.* Raytracing with MARX: X-ray observatory design, calibration, and support. *Proc. SPIE* **8443**, 84431A (2012).
35. Hong, J. *et al.* NuSTAR hard X-ray survey of the galactic center region. II. X-ray point Sources. *Astrophys. J.* **825**, 132 (2016).
36. Sakano, M. *et al.* Unusual X-ray transients in the Galactic Centre. *Mon. Not. R. Astron. Soc.* **357**, 1211–1218 (2005).
37. Ponti, G. *et al.* A powerful flare from Sgr A* confirms the synchrotron nature of the X-ray emission. *Mon. Not. R. Astron. Soc.* **468**, 2447–2468 (2017).
38. Takata, J., Cheng, K. S. & Taam, R. E. X-ray and gamma-ray emissions from different evolutionary stage of rotation powered millisecond pulsars. *Astrophys. J.* **745**, 100 (2012).
39. Kaspi, V. M. & Beloborodov, A. M. Magnetars. *Annu. Rev. Astron. Astrophys.* **55**, 261–301 (2017).
40. Rockefeller, G., Fryer, C. L., Baganoff, F. K. & Melia, F. The X-ray ridge surrounding Sagittarius A* at the Galactic Center. *Astrophys. J.* **635**, L141–L144 (2005).
41. Overbeck, J. W. Small-angle scattering of celestial X-rays by interstellar grains. *Astrophys. J.* **141**, 864–866 (1965).
42. Jin, C. *et al.* Probing the interstellar dust towards the Galactic Centre: dust-scattering halo around AX J1745.6–2901. *Mon. Not. R. Astron. Soc.* **468**, 2532–2551 (2017).
43. Corrales, L. R. *et al.* The Chandra dust-scattering halo of Galactic Center transient Swift J174540.7–290015. *Astrophys. J.* **839**, 76 (2017).
44. Corrales, L. *et al.* The dust-scattering component of X-ray extinction: effects on continuum fitting and high-resolution absorption edge structure. *Mon. Not. R. Astron. Soc.* **458**, 1345–1351 (2016).
45. Bradley, C. K. The spectrum of the black hole X-ray nova V404 Cygni in quiescence as measured by XMM-Newton. *Astrophys. J.* **667**, 427–432 (2007).
46. Scargle, J. D., Norris, J. P., Jackson, B. & Chiang, J. Studies in astronomical time series analysis. VI. Bayesian block representations. *Astrophys. J.* **764**, 167 (2013).
47. Overbeck, J. W. Small-angle scattering of celestial X-rays by interstellar grains. *Astrophys. J.* **141**, 864–866 (1965).
48. Bogdanov, S. *et al.* Chandra X-ray observations of the 12 millisecond pulsars in the globular cluster M28. *Astrophys. J.* **730**, 81 (2011).
49. Schödel, R. *et al.* The distribution of stars around the Milky Way's central black hole. II. Diffuse light from sub-giants and dwarfs. *Astron. Astrophys.* <https://doi.org/10.1051/0004-6361/201730452> (2017).
50. Armas Padilla, M. *et al.* Swift J1357.2–0933: the faintest black hole? *Mon. Not. R. Astron. Soc.* **444**, 902–905 (2014).
51. Casares, J. *et al.* A Be-type star with a black-hole companion. *Nature* **505**, 378–381 (2014).



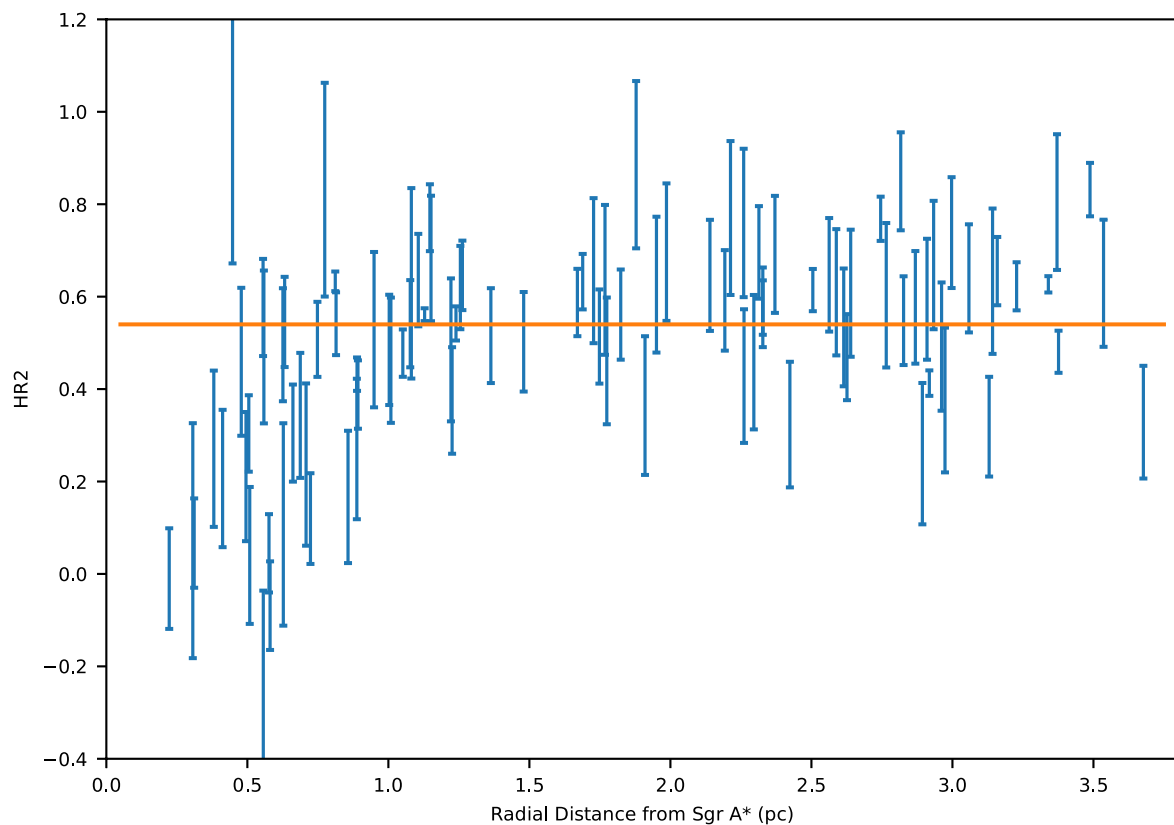
Extended Data Figure 1 | Example Chandra spectra of thermal and non-thermal X-ray sources. **a**, CXO J174541.02–290017.6 (HR2 > 0.3). **b**, CXO J174539.48–290045.8 (HR2 < 0.3). The hard (thermal) source spectrum, which exhibits a broad emission line at $E = 6\text{--}7$ keV, is fitted

by a partially covered APEC model, whereas the soft (non-thermal) source spectrum is fitted by an absorbed power-law model. The error bars represent 1σ statistical uncertainties. The bottom panels show residuals (data minus model) in terms of 1σ significance.



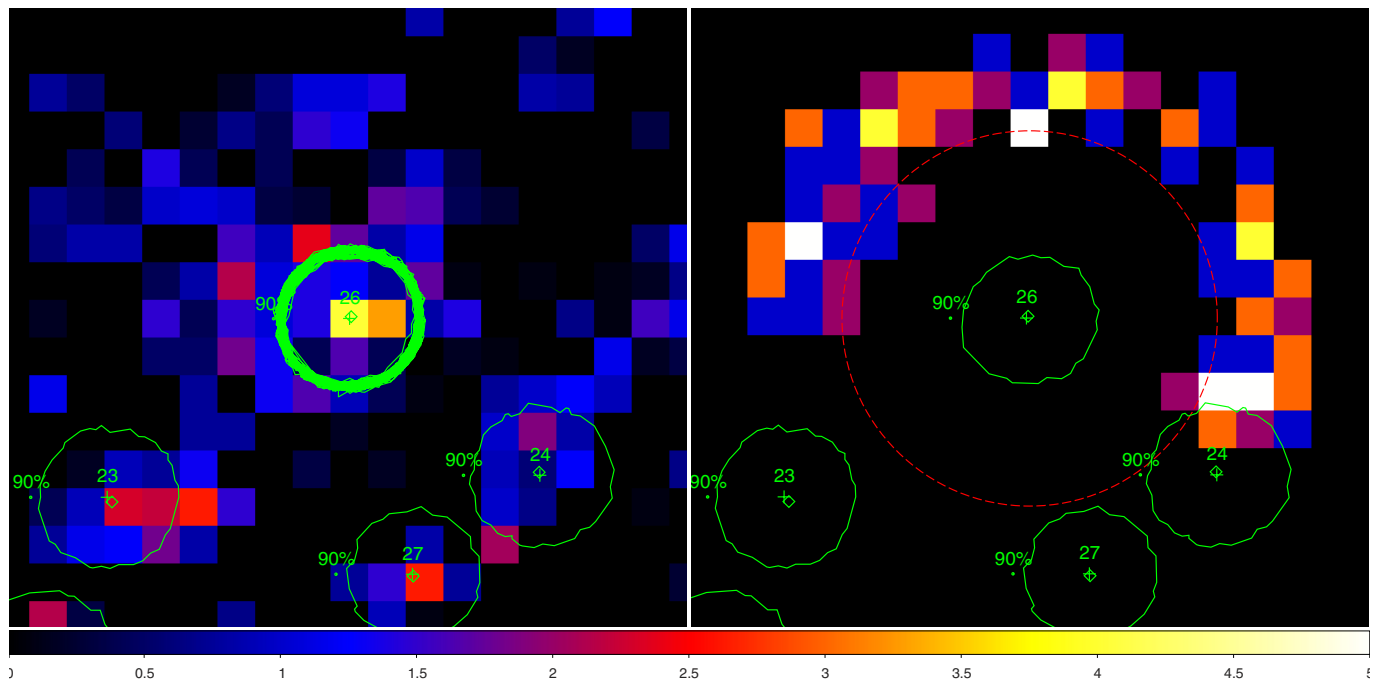
Extended Data Figure 2 | A cumulative X-ray flux distribution of the qBH-LMXB candidates within 1 pc. A logN–logS histogram of the 12 $\text{HR2} < 0.3$ sources (assuming all are qBH-LMXBs) is plotted, where S denotes an absorbed 2–8-keV flux. The best-fitting power-law model is

shown as a solid blue line. An extrapolation of this line to the minimum observed flux (1.9×10^{-16} erg cm^{-2} s^{-1} , after correcting for higher X-ray extinction in the Galactic Centre) of the local qBH-LMXB population⁵⁰ is used to estimate the total number of qBH-LMXBs in the central parsec.



Extended Data Figure 3 | Hardness ratio for X-ray point sources with $C \geq 100$ as a function of the projected radial distance from Sgr A*. The horizontal line denotes the best fit to a radially constant hardness ratio ($HR2 = 0.53$), which yields a poor fit with $\chi^2_\nu = 3.0$ for 89 d.o.f. We took

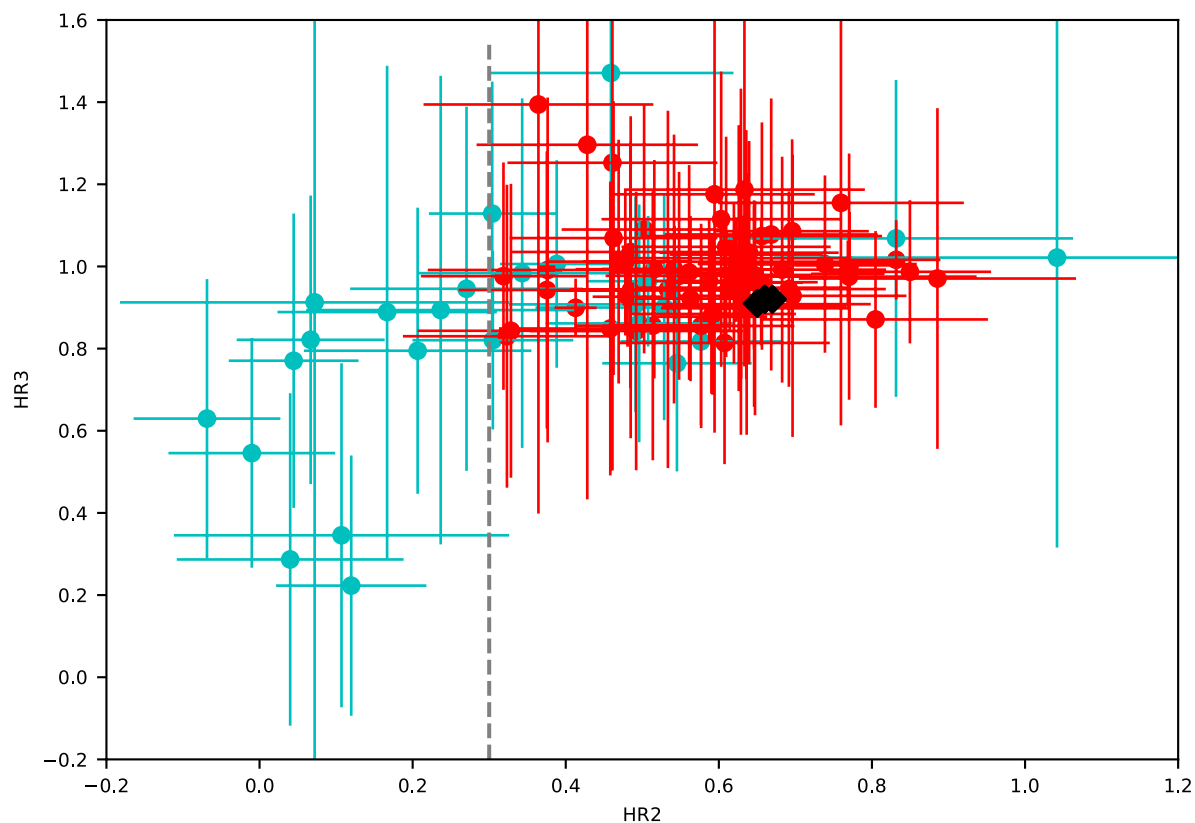
into account statistical and systematic errors in the HR2 data; the error bars in the figure indicate 1σ statistical uncertainties. The hypothesis of spatial constancy of HR2 is thus rejected with $P = 1.1 \times 10^{-19}$. The bifurcation in HR2 at $r \lesssim 1$ pc is clearly visible.



Extended Data Figure 4 | Chandra images of the soft source CXO J174540.79–290024.5 and its extracted background events.

a, Neighborhood image around the source (labelled 26, with multiple green circles near the centre indicating source extraction regions for different observations) along with other Chandra catalogue sources indicated by numbers 23, 24 and 27. **b**, Background image around the

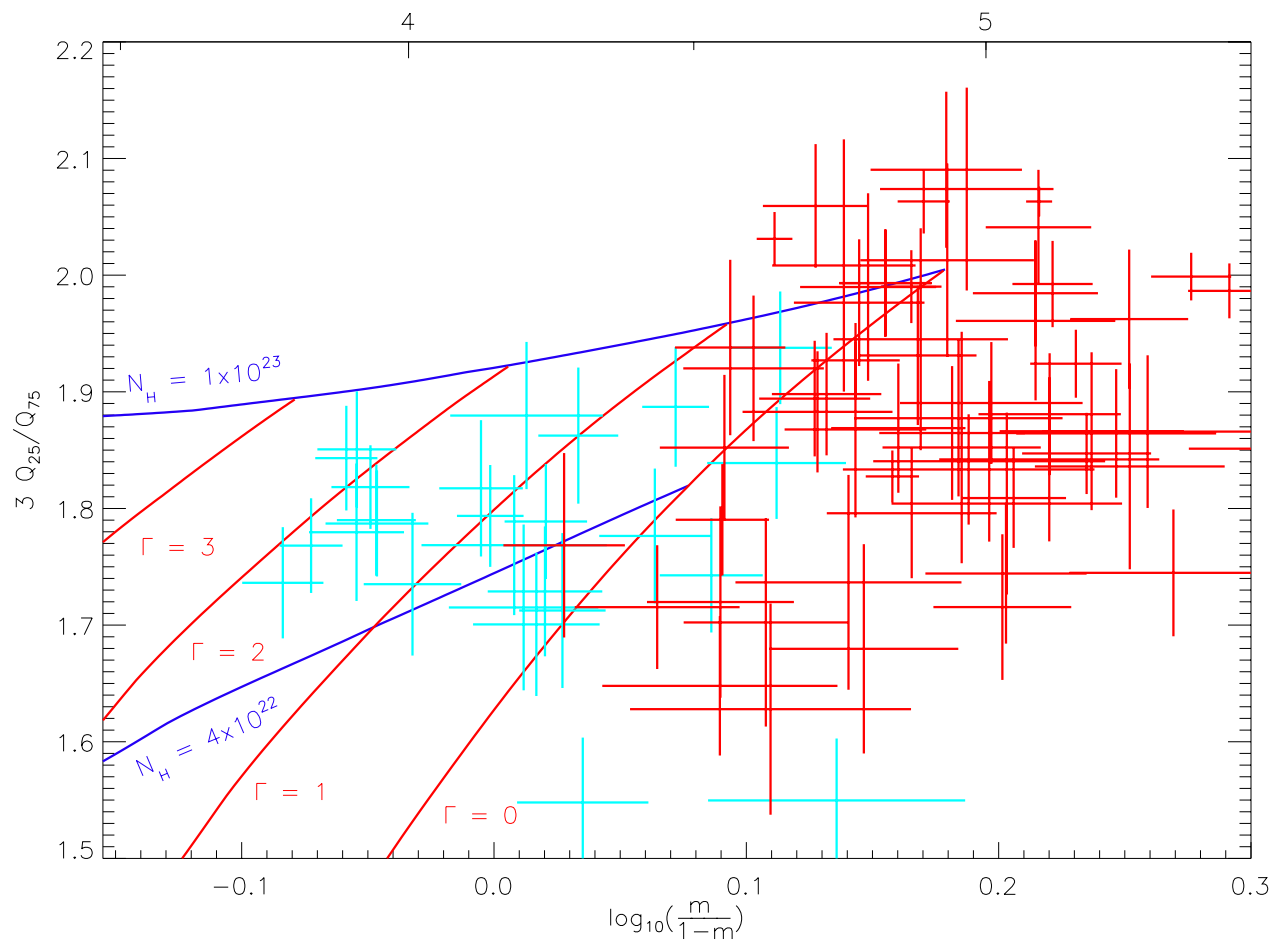
source (labelled 26 and located near the centre with a green circle) from ObsID 11843. The red dashed circle indicates the inner circle of the background annulus ($r = 2.5''$). Note that background was properly extracted from an annulus of $r \approx 2.5'' - 4''$ by excluding the nearby Chandra sources indicated by green circles labelled 23, 24 and 27.



Extended Data Figure 5 | Colour-colour diagram of X-ray sources with $C \geq 100$ within radial projected distances from Sgr A*. HR2 and HR3 hardness data are plotted for sources at $r < 1$ pc (cyan) and $1 \text{ pc} < r < 3.8$ pc (red). The error bars represent 1σ statistical uncertainties. There is no clear evidence of HR3–HR2 correlation, a surrogate for column-density- or scattering-related effects. The vertical dotted grey line

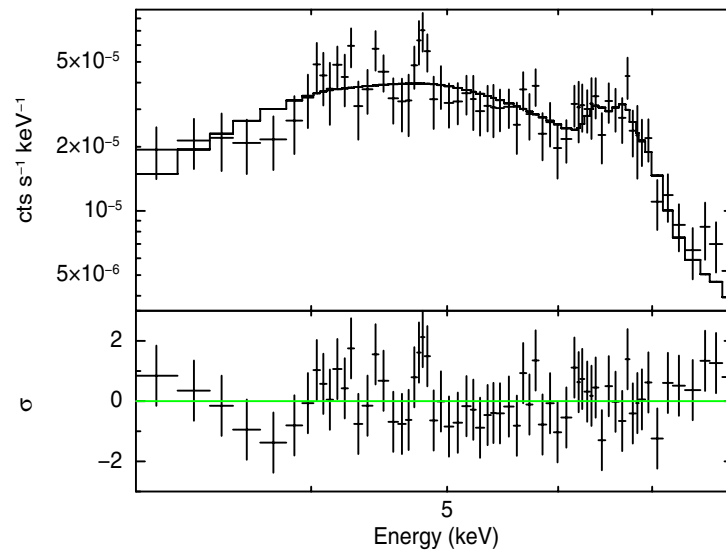
HR2

indicates our dividing line at $\text{HR2} = 0.3$ between the soft and hard sources. For reference, the calculated HR2 and HR3 for four local IPs, whose spectra were extrapolated to the distance of Sgr A*, are shown as black diamonds. They cluster at high HR2, as expected. Only three of the four IP diamonds are visible because two of them overlap.



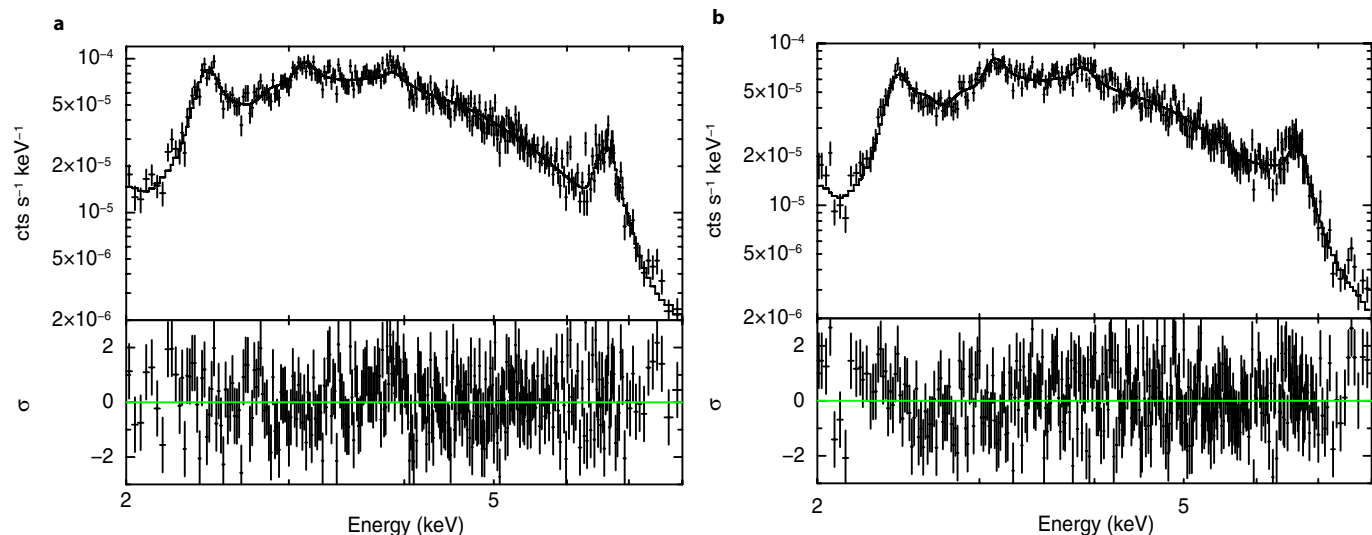
Extended Data Figure 6 | Median photon energy in the spectrum versus hardness ratio of X-ray sources with $C \geq 100$. Sources for $r < 1$ pc and $r > 1$ pc are indicated in cyan and red, respectively. The median $m = Q_{50}$ and the ratio of two quartiles Q_{25}/Q_{75} are calculated for each source¹⁴. The

energy scale across the top shows the median energy values ($E_{50\%}$). The blue and red lines correspond to hydrogen column density N_H and power-law photon index Γ grids, respectively. The error bars represent 1σ statistical uncertainties.



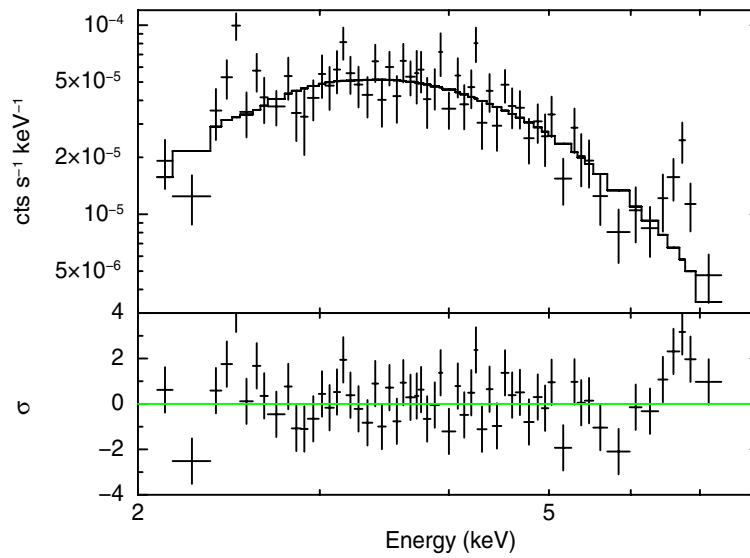
Extended Data Figure 7 | Stacked Chandra ACIS-I spectrum of the 14 thermal X-ray sources with net counts $C \geq 100$ within $r < 1$ pc. The spectrum is well fitted with a partially covered thermal APEC model with a Gaussian line at 6.4 keV ($\chi^2_\nu = 0.80$ for 58 d.o.f.). The error bars

represent 1σ statistical uncertainties. The bottom panels show residuals (data minus model) in terms of 1σ significance. The best-fit parameters are shown in Extended Data Table 2.



Extended Data Figure 8 | Stacked background Chandra ACIS-I spectra of the sources with net counts $C \geq 100$ within $r < 1$ pc. a, 12 HR2 < 0.3 sources. b, 14 HR2 > 0.3 sources. Several prominent emission lines are present from low- Z elements (below 4 keV) and Fe (at 6–7 keV). Both background spectra are well fitted with an absorbed two-temperature thermal APEC model: $\chi^2_\nu = 1.21$ (291 d.o.f.; soft source background) and $\chi^2_\nu = 1.26$ (320 d.o.f.; hard source background). The best-fit temperatures

are consistent between the two background spectra: $kT_1 = 0.93^{+0.10}_{-0.04}$ keV and $kT_2 = 5.9^{+1.5}_{-2.9}$ keV (for the soft source background) and $kT_1 = 1.18^{+0.05}_{-0.04}$ keV and $kT_2 = 6.6^{+1.5}_{-1.1}$ keV (for the hard source background). The error bars represent 1σ statistical uncertainties. The bottom panels show residuals (data minus model) in terms of 1σ significance.



Extended Data Figure 9 | Simulated Chandra ACIS-I spectrum using a two-temperature thermal APEC model. The input parameters are $kT_1 = 1.0$ keV and $Z_1 = 2.5$ (for the lower-temperature component) and $kT_2 = 5.0$ keV and $Z_2 = 0.6$ (for the higher-temperature component). The spectrum was poorly fit with an absorbed power-law model yielding

$\chi^2_\nu = 1.65$ (56 d.o.f.; hard source background) and showing large residuals at 2–3 keV and 6–7 keV. The best-fit photon index ($\Gamma = 3.1^{+0.4}_{-0.4}$) is significantly softer than that of the stacked soft source spectrum in Fig. 2b ($\Gamma = 1.9^{+0.3}_{-0.3}$).

Extended Data Table 1 | The 12 non-thermal (soft) X-ray sources with $C \geq 100$ in the central parsec

Source name	Projected Distance from Sgr A* [pc]	Net counts (2-8 keV)	Long term X-ray variability	Wavdetect Significance (σ)
174539.87-290034.2	0.23	261	TRUE	15.6
174540.38-290033.5	0.32	100	FALSE	4.6
174540.40-290024.1	0.32	301	FALSE	15.2
174540.45-290036.3	0.43	171	TRUE	8.1
174540.79-290024.5	0.52	193	TRUE	24.2
174539.40-290040.9	0.60	309	TRUE	21.7
174540.95-290031.2	0.60	274	FALSE	14.7
174541.03-290026.8	0.65	122	FALSE	15.4
174540.63-290013.4	0.73	130	FALSE	6.7
174539.48-290045.8	0.75	268	TRUE	32.6
174540.37-290049.9	0.88	136	FALSE	10.8
174539.28-290049.1	0.92	136	TRUE	8.4

The last column indicates the source detection significance calculated by *wavdetect*, a wavelet algorithm used widely for Chandra analysis. The sources have $HR2 < 0.3$.

Extended Data Table 2 | Best-fitting parameters from fitting the stacked Chandra spectra

Parameters	Hard Sources ($HR2 > 0.3$)	Soft Sources ($HR2 < 0.3$)
Model	fgcdust*tbabs*pcfabs*(apec+gauss)	fgcdust*tbabs*powerlaw
N_H [10^{23}cm^{-2}] (Galactic)	$1.1^{+1.3}_{-0.7}$	$0.8^{+0.2}_{-0.2}$
N_H [10^{23}cm^{-2}] (Partial covering)	20^{+68}_{-16}	—
Covering fraction	$0.9^{+0.1}_{-0.8}$	—
kT [keV]	$7.3^{+3.0}_{-1.3}$	—
A_{Fe}	$0.6^{+0.3}_{-0.2}$	—
Γ	—	$1.9^{+0.3}_{-0.3}$
2-8 keV flux [$10^{-14}\text{erg cm}^{-2} \text{ s}^{-1}$] (absorbed)	$8.7^{+0.3}_{-0.3}$	$5.9^{+0.1}_{-1.8}$
χ^2_ν (dof)	0.80 (58)	0.81 (51)

The errors indicate 1σ statistical uncertainties. The spectral models are a partially covered APEC model with a Gaussian line and an absorbed power-law model for the 14 hard ($HR2 > 0.3$) and the 12 soft sources ($HR2 < 0.3$), respectively. The line energy and width of the Gaussian line component are fixed to 6.4 keV and 0.01 keV to account for the neutral Fe-K fluorescence line. The parameters of the partial covering absorption component are poorly constrained owing to spectral fitting in the narrow 2–8-keV energy band.

Quantized Majorana conductance

Hao Zhang^{1*}, Chun-Xiao Liu^{2*}, Sasa Gazibegovic^{3*}, Di Xu¹, John A. Logan⁴, Guanzhong Wang¹, Nick van Loo¹, Jouri D. S. Bommer¹, Michiel W. A. de Moor¹, Diana Car³, Roy L. M. Op het Veld³, Petrus J. van Veldhoven³, Sebastian Koelling³, Marcel A. Verheijen^{3,5}, Mihir Pendharkar⁶, Daniel J. Pennachio⁴, Borzoyeh Shojaei^{4,7}, Joon Sue Lee⁷, Chris J. Palmstrøm^{4,6,7}, Erik P. A. M. Bakkers³, S. Das Sarma² & Leo P. Kouwenhoven^{1,8}

Majorana zero-modes—a type of localized quasiparticle—hold great promise for topological quantum computing¹. Tunnelling spectroscopy in electrical transport is the primary tool for identifying the presence of Majorana zero-modes, for instance as a zero-bias peak in differential conductance². The height of the Majorana zero-bias peak is predicted to be quantized at the universal conductance value of $2e^2/h$ at zero temperature³ (where e is the charge of an electron and h is the Planck constant), as a direct consequence of the famous Majorana symmetry in which a particle is its own antiparticle. The Majorana symmetry protects the quantization against disorder, interactions and variations in the tunnel coupling^{3–5}. Previous experiments⁶, however, have mostly shown zero-bias peaks much smaller than $2e^2/h$, with a recent observation⁷ of a peak height close to $2e^2/h$. Here we report a quantized conductance plateau at $2e^2/h$ in the zero-bias conductance measured in indium antimonide semiconductor nanowires covered with an aluminium superconducting shell. The height of our zero-bias peak remains constant despite changing parameters such as the magnetic field and tunnel coupling, indicating that it is a quantized conductance plateau. We distinguish this quantized Majorana peak from possible non-Majorana origins by investigating its robustness to electric and magnetic fields as well as its temperature dependence. The observation of a quantized conductance plateau strongly supports the existence of Majorana zero-modes in the system, consequently paving the way for future braiding experiments that could lead to topological quantum computing.

A semiconductor nanowire coupled to a superconductor can be tuned into a topological superconductor with two Majorana zero-modes localized at the wire ends^{1,8,9}. Tunnelling into a Majorana mode will show a zero-energy state in the tunnelling density-of-states, that is, a zero-bias peak (ZBP) in the differential conductance (dI/dV)^{2,6}. This tunnelling process is an ‘Andreev reflection’, in which an incoming electron is reflected as a hole. Particle–hole symmetry dictates that the zero-energy tunnelling amplitudes of electrons and holes are equal, resulting in a perfect resonant transmission with a ZBP height quantized at $2e^2/h$ (refs 3, 4, 10), irrespective of the precise tunnelling strength^{3–5}. The Majorana nature of this perfect Andreev reflection is a direct result of the well-known Majorana symmetry property ‘particle equals antiparticle’^{11,12}.

This predicted robust conductance quantization has not yet been observed^{2,6,7,13,14}. Instead, most of the ZBPs have a height considerably less than $2e^2/h$. This discrepancy was first explained by thermal averaging^{15–18}, but that explanation does not hold when the peak width exceeds the thermal broadening (about $3.5k_B T$)^{13,14}. In that case, other averaging mechanisms, such as dissipation¹⁹, have been invoked. The main source of dissipation is a finite quasiparticle density-of-states

within the superconducting gap, often referred to as a ‘soft gap’. Substantial advances have been achieved in ‘hardening’ the gap by improving the quality of materials, eliminating disorder and interface roughness^{20,21}, and better control during device processing^{22,23}, all guided by a more detailed theoretical understanding²⁴. We have recently solved all these dissipation and disorder issues²¹, and here we report the resulting improvements in electrical transport leading to the elusive quantization of the Majorana ZBP.

Figure 1a shows a micrograph of a fabricated device and schematics of the measurement set-up. An InSb nanowire (grey) is partially covered (two out of six facets) by a thin superconducting aluminium shell (green)²¹. The ‘tunnel-gates’ (coral red) are used to induce a tunnel barrier in the non-covered segment between the left electrical contact (yellow) and the Al shell. The right contact is used to drain the current to ground. The chemical potential in the segment covered with Al can be tuned by applying voltages to the two long ‘super-gates’ (purple).

Transport spectroscopy is shown in Fig. 1b, which displays dI/dV as a function of voltage bias V and magnetic field B (aligned with the nanowire axis), while fixed voltages are applied to the tunnel- and super-gates. As B increases, two levels detach from the gap edge (at about 0.2 meV), merge at zero bias and form a robust ZBP. This is consistent with the Majorana theory: a ZBP is formed after the Zeeman energy closes the trivial superconducting gap and re-opens a topological gap^{8,9}. The gap re-opening is not visible in a measurement of the local density-of-states because the tunnel coupling to these bulk states is small²⁵. Moreover, the finite length (about 1.2 μm) of the proximitized segment (that is, the part that is superconducting because of the proximity effect from the superconducting Al coating) results in discrete energy states, turning the trivial-to-topological phase transition into a smooth crossover²⁶. Figure 1c shows two line-cuts from Fig. 1b extracted at 0 T and 0.88 T. Importantly, the height of the ZBP reaches the quantized value of $2e^2/h$. The line-cut at zero bias in the lower panel of Fig. 1b shows that the ZBP height remains close to $2e^2/h$ over a sizable range in B field (0.75–0.92 T). Beyond this range, the height drops, most probably because of a closure of the superconducting gap in the bulk Al shell.

We note that the sub-gap conductance at $B = 0$ (black curve, left panel, Fig. 1c) is not completely suppressed down to zero, reminiscent of a soft gap. In this case, however, this finite sub-gap conductance does not reflect any finite sub-gap density-of-states in the proximitized wire. It arises from Andreev reflection (that is, transport by dissipationless Cooper pairs) due to a high tunnelling transmission, which is evident from the above-gap conductance (dI/dV for $V > 0.2$ mV) being larger than e^2/h . As this softness does not result from dissipation, the Majorana peak height should still reach the quantized value²⁷. In Extended Data Fig. 1, we show that this device tuned into a low-transmission regime,

¹QuTech and Kavli Institute of NanoScience, Delft University of Technology, 2600 GA Delft, The Netherlands. ²Condensed Matter Theory Center and Joint Quantum Institute, Department of Physics, University of Maryland, College Park, Maryland 20742, USA. ³Department of Applied Physics, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands. ⁴Materials Engineering, University of California Santa Barbara, Santa Barbara, California 93106, USA. ⁵Philips Innovation Services Eindhoven, High Tech Campus 11, 5656AE Eindhoven, The Netherlands. ⁶Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, California 93106, USA. ⁷California NanoSystems Institute, University of California Santa Barbara, Santa Barbara, California 93106, USA. ⁸Microsoft Station Q Delft, 2600 GA Delft, The Netherlands.

*These authors contributed equally to this work.

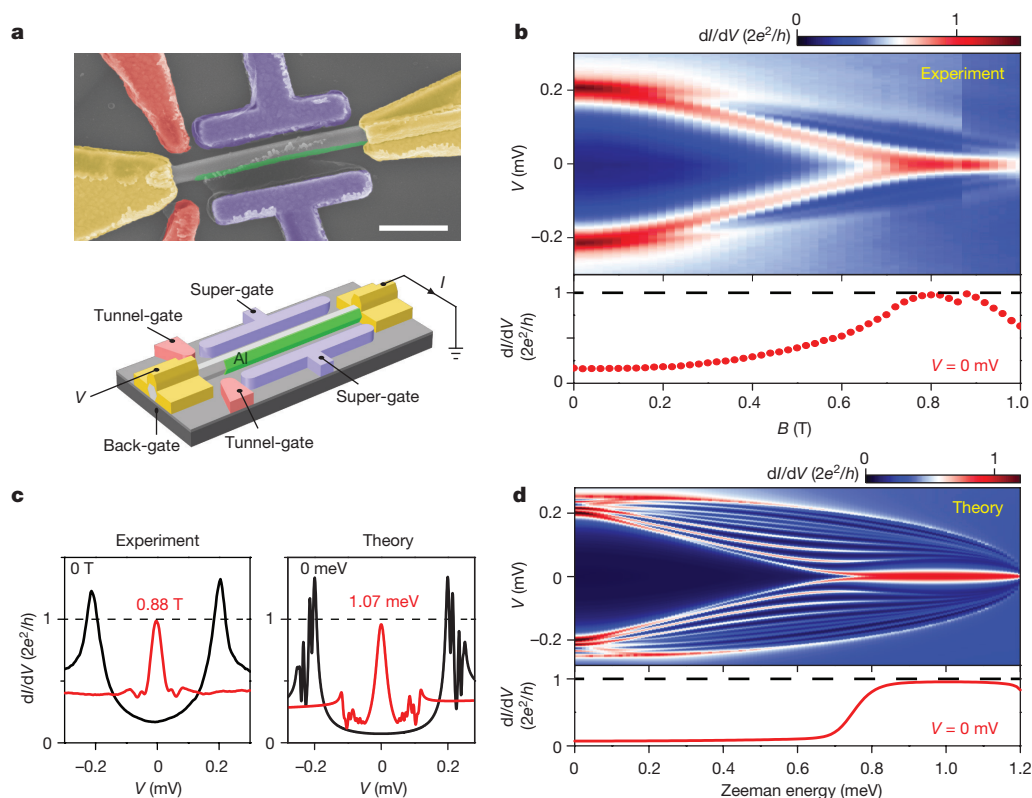


Figure 1 | Quantized Majorana zero-bias peak. **a**, False-colour scanning electron micrograph of device A (upper panel) and its schematics (lower panel). Side gates and contacts are Cr/Au (10 nm/100 nm). The Al shell thickness is approximately 10 nm. The substrate is p-doped Si, acting as a global back-gate, covered by 285 nm SiO₂. The two tunnel-gates are shorted externally, as are the two super-gates. Scale bar, 500 nm. **b**, Magnetic field dependence of the quantized ZBP in device A with the zero-bias line-cut in the lower panel. Magnetic field direction is aligned with the nanowire axis for all measurements. Super-gate (tunnel-gate) voltage is fixed at -6.5 V (-7.7 V), while the back-gate is

kept grounded. Temperature is 20 mK unless specified. **c**, Comparison between experiment and theory. Left (right) panel shows the vertical line-cuts from **b** (**d**) at 0 T and 0.88 T (1.07 meV). **d**, Majorana simulation of device A, assuming chemical potential $\mu = 0.3$ meV, tunnel barrier length ($L_{TG} = 10$ nm), with height $E_{TG} = 8$ meV, and the superconductor–semiconductor coupling is 0.6 meV. See Methods for further information. A small dissipation broadening term (about 30 mK) is introduced for all simulations to account for the averaging effect from finite temperature and small lock-in excitation voltage (8 μ V).

where dI/dV does reflect the density-of-states, displays a hard gap (also shown in Extended Data Fig. 4, where the gap remains hard in a magnetic field). For further understanding, we use experimental parameters in a theoretical Majorana nanowire model²⁸ (see Methods for more information). Figure 1d shows a simulation with two line-cuts shown in Fig. 1c (right panel). Besides the ZBP, other discrete sub-gap states are visible, which are due to the finite wire length. Such discrete lines are only faintly resolved in the experimental panels of Fig. 1b. Overall, we find good qualitative agreement between the experimental and simulation panels in Fig. 1b and d. An exact quantitative agreement is not feasible, as the precise experimental values for the parameters going into the theory (for example, chemical potential, tunnel coupling, Zeeman splitting or spin–orbit coupling) are unknown for our hybrid wire–superconductor structure.

Next, we fix B at 0.8 T and investigate the robustness of the quantized ZBP against variations in transmission by varying the voltage on the tunnel-gate. Figure 2a shows dI/dV while varying V and tunnel-gate voltage. Figure 2b shows that the ZBP height remains close to the quantized value. Importantly, the above-gap conductance measured at $|V| = 0.2$ meV varies by more than 50% (Fig. 2c and d), implying that the transmission is changing considerably over this range while the ZBP remains quantized. The minor conductance switches in Fig. 2a–c are due to unstable jumps of trapped charges in the surroundings.

Figure 2d (red curves) shows several line-cuts of the quantized ZBP. The extracted height and width are plotted in Fig. 2e (upper panel) as a function of above-gap conductance $G_N = T \times e^2/h$ where T is the transmission probability for a spin-resolved channel. Although the ZBP

width does change with G_N , the quantized height remains unaffected. Note that the ZBP width ranges from about 50 μ eV to about 100 μ eV, which is significantly wider than the thermal width of approximately 6 μ eV at 20 mK. The ZBP width is thus broadened by tunnel coupling, instead of thermal broadening, fulfilling a necessary condition to observe a quantized Majorana peak. In Extended Data Fig. 2, we show that in the low-transmission regime in which thermal broadening dominates over tunnel broadening, the ZBP height drops below $2e^2/h$ (as explained in refs 15–18). The robustness of the ZBP quantization to a variation in the tunnel barrier is an important finding of our work.

A more negative tunnel-gate voltage (< -8 V) eventually splits the ZBP, which may be explained by an overlapping of the two localized Majorana wavefunctions from the two wire ends. The tunnel-gate not only tunes the transmission of the barrier but also influences the potential profile in the proximitized wire part near the tunnel barrier. A more negative gate voltage effectively pushes the nearby Majorana mode away, towards the remote Majorana on the other end of the wire, thus reducing the length of the effective topological wire. This leads to the wavefunction overlap between the two Majorana modes, causing the ZBP to split¹⁶ (black curves in Fig. 2d). This splitting is also captured in our simulations shown in Fig. 2f, where we have checked that the splitting originates from Majorana wavefunction overlap. Note that the simulated ZBP height (red curve in middle panel in Fig. 2f) remains close to the $2e^2/h$ plateau over a large range, whereas the above-gap conductance (black curve in lower panel in Fig. 2f) changes substantially. Also, the height and width dependence in the simulation is in qualitative agreement with our experimental observation (Fig. 2e).

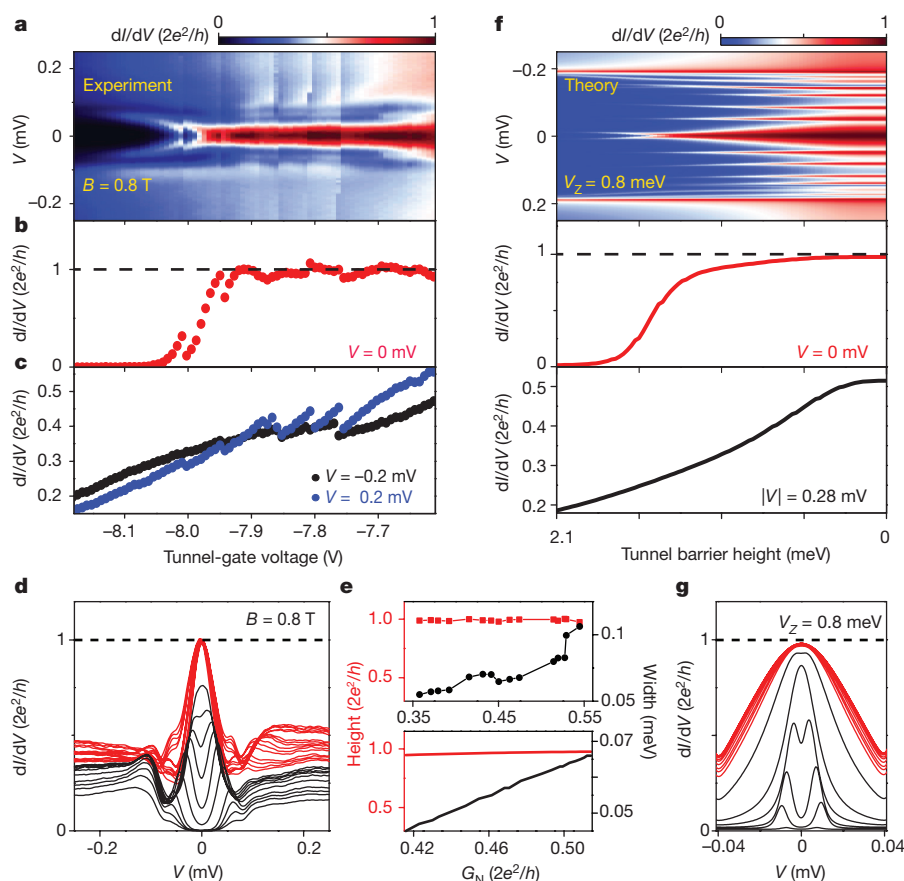


Figure 2 | Quantized Majorana conductance plateau. **a**, Tunnel-gate dependence of the quantized ZBP at $B = 0.8$ T. Super-gate (back-gate) voltage is fixed at -6.5 V (0 V). **b**, **c**, Horizontal line-cuts from **a**, showing zero-bias conductance and above-gap conductance, respectively. The zero-bias conductance shows a quantized plateau. **d**, Several vertical line-cuts from **a**, showing ZBPs with quantized height (red curves). For the black curves, the zero-bias conductance drops below the quantized value owing to peak splitting. **e**, (Upper panel) ZBP height (red squares) and width (black dots) extracted from **d** (red curves), as a function of

above-gap conductance (G_N). The width is defined by the bias voltage at which $dI/dV = e^2/h$. (Lower panel) ZBP height and width extracted from several simulation curves in **f**. **f**, Majorana simulation of the tunnel-gate dependence. We set the Zeeman field $V_z = 0.8$ meV and chemical potential $\mu = 0.6$ meV, such that the nanowire is in the topological regime. From left to right, the barrier width decreases linearly from 175 nm to 0 nm, as the barrier height decreases from 2.1 meV to 0. **g**, Vertical line-cuts from **f** show the quantized ZBP (red) and split peaks (black).

To complete the comparison, we show in Fig. 2g the simulated line-cuts of several quantized ZBPs (red curves) and split peaks (black curves), consistent with the experimental data in Fig. 2d.

Pushing Majorana modes towards each other is one mechanism for splitting. Another way is by changing the chemical potential through the transition from a topological to a trivial phase^{8,9}—the quantum phase transition from the trivial to the topological phase can equivalently be caused by tuning either the Zeeman energy (that is, the magnetic field) or the chemical potential. Splitting at the phase transition occurs because the Majorana wavefunctions start to spread out over the entire wire length. For long wires, the transition is abrupt, whereas in shorter wires a smooth transition is expected²⁶. We investigate the dependence of the quantized ZBP on chemical potential by varying the voltage on the super-gate. Figure 3a shows a nearly quantized ZBP that remains non-split over a large range in the super-gate voltage. More positive voltage applied to the super-gates corresponds to a higher chemical potential, and eventually we find a ZBP splitting (around -5 V or more positive) and consequently a suppression of the zero-bias conductance below the quantized value. Although the relation between the gate voltage and chemical potential is unknown in our devices, this splitting suggests a transition to the trivial phase caused by a tuning of the chemical potential induced by the changing super-gate voltage.

In a lower B field and different gate settings (Fig. 3b), the splitting of the quantized ZBP shows oscillatory behaviour as a function of the super-gate voltage. The five line-cuts on the right panel highlight this

back-and-forth behaviour between quantized and suppressed ZBPs. Notably, the ZBP height comes back up to the quantized value and does not cross through it.

We find similar behaviour in the theoretical simulations of Fig. 3c. In these simulations, we have confirmed that for the chosen parameters, the Majorana wavefunctions oscillate in their overlap, thus giving rise to the back-and-forth behaviour of quantized and split ZBPs²⁹. In the experiment, it may also be that non-homogeneity, possibly somewhere in the middle of the wire, causes overlap of Majorana wavefunctions. Again, we note that the conversion from gate voltage to chemical potential is unknown, preventing a direct quantitative comparison between experiment and simulation.

To demonstrate the reproducibility of ZBP quantization, we show in Fig. 4a the quantized ZBP data from a second device. In this second device, the length of the proximitized section is about $0.9\ \mu\text{m}$, which is about $0.3\ \mu\text{m}$ shorter than in the previous device. The quantized ZBP plateau is indicated by the region between the two green dashed lines in Fig. 4b (red curve). This second device allows transmission of more than one channel through the tunnel barrier, which we deduce from the above-gap conductance value (Fig. 4b, lower panel, black curve) exceeding e^2/h for tunnel-gate voltages higher than about -0.55 V. Correspondingly, the zero-bias conductance can now exceed $2e^2/h$ (Fig. 4b, middle panel) for such an open tunnel barrier⁵. Tunnelling through the second channel in the barrier region results in an additional background conductance, thus leading to the zero-bias conductance

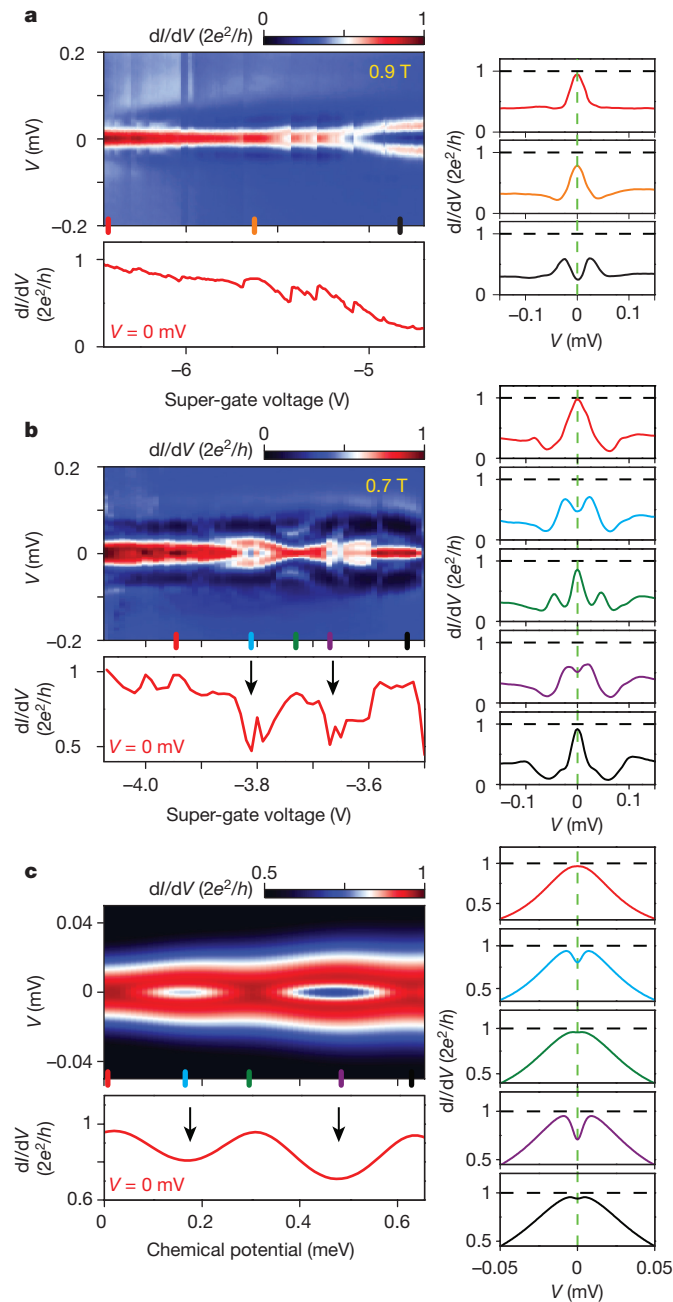


Figure 3 | Majorana peak splitting. **a**, Super-gate dependence of the quantized ZBP in device A at 0.9 T. As the super-gate increases the chemical potential, the ZBP height is nearly quantized before it splits. The tunnel-gate voltage is adjusted simultaneously when sweeping the super-gate voltage, to compensate for the cross coupling and keep the transmission roughly constant. Lower panel shows the zero-bias line-cut, and the right panels show vertical line-cuts at gate voltages indicated by the corresponding colour bars. Switches in the colour maps are due to charge jumps in the gate dielectric. **b**, Oscillatory behaviour of the ZBP splitting, where the two black arrows point at the peak splitting regions. **c**, Simulation also shows oscillatory splitting as a function of chemical potential. The Zeeman field is fixed at $V_z = 1$ meV.

rising above $2e^2/h$. We find, however, from a rough estimate of this background contribution that the net ZBP height (above background) never exceeds $2e^2/h$, consistent with Majorana theory⁵.

We next fix the B field and study temperature dependence. Figure 4c shows a line-cut of this quantized ZBP from Fig. 4a. First, the base temperature trace in Fig. 4c (red data points) fits well to a Lorentzian line-shape with a 20 mK thermal broadening, expected for Majoranas³⁰

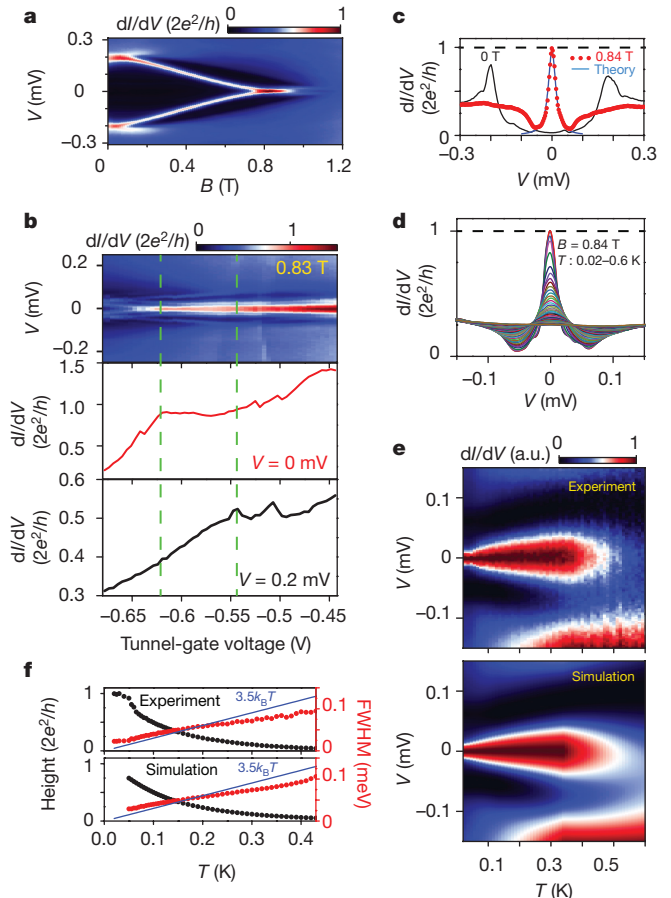


Figure 4 | Quantized Majorana plateau reproduced, and temperature dependence. **a**, Magnetic field dependence of the quantized ZBP in device B. **b**, Tunnel-gate dependence of the ZBP at 0.83 T. The two lower panels are the horizontal line-cuts at bias voltage, V , of 0 mV and 0.2 mV. The two dashed green lines indicate the plateau region of the zero-bias conductance. **c**, Vertical line-cuts from **a** at 0 T and 0.84 T. The blue line is a Lorentzian fit with a tunnel coupling $\Gamma = 13.7 \mu\text{eV}$ and temperature of 20 mK. **d**, Temperature dependence of this quantized ZBP while the temperature increases from 20 mK to 600 mK in steps of 10 mK. **e**, Colour plot of the temperature dependence in the upper panel with the simulation in the lower panel. At each temperature, the conductance is renormalized by setting the minimum to 0 and maximum to 1, for clarity. a.u., arbitrary units. **f**, Extracted ZBP height and FWHM as a function of temperature from **e**. Upper panel is the experiment; lower panel is the simulation with no fitting parameters.

as well as for any type of resonant transmission. The ZBP temperature dependence is shown in line traces in Fig. 4d and in colour scale in Fig. 4e (with the corresponding simulation in the lower panel of Fig. 4e). Figure 4f shows the extracted ZBP height and ZBP width (full-width at half-maximum, FWHM) from both the experimental and simulated traces. At low temperatures, the ZBP width (red data points) exceeds the thermal width defined as $3.5k_B T$ (blue line). In agreement with theory³¹, the ZBP height (black data points) reaches and saturates at $2e^2/h$ when the FWHM exceeds $3.5k_B T$. For higher temperatures, thermal averaging starts to suppress the ZBP height below the quantized value. The simulated data are calculated by a convolution of the derivative of the Fermi distribution function and the dI/dV trace at a base temperature of 20 mK. This procedure of incorporating thermal effects holds if the temperature of the calculated dI/dV curve is significantly larger than base temperature (which can then be assumed to be the effective zero-temperature conductance value). We find excellent agreement between experiment and simulation for $T > 50$ mK (Fig. 4f). See Extended Data Fig. 3 for detailed temperature dependence.

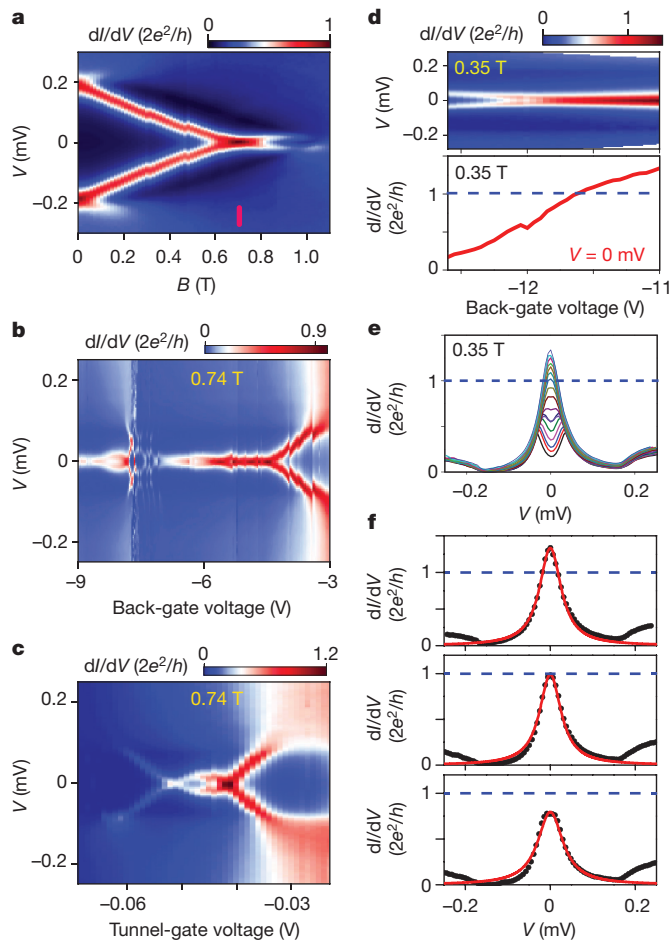


Figure 5 | Trivial zero-bias peaks from Andreev bound states.

a, Magnetic field dependence of a trivial ZBP in device C. The peak height reaches $2e^2/h$ at 0.7 T (red bar). **b**, Back-gate dependence of this ZBP, where the peak remains non-split for a sizable range of gate voltage. The peak height varies and is generally below $2e^2/h$. **c**, Tunnel-gate dependence of this ZBP, which is a result of level crossing. **d**, Back-gate dependence of the ZBP at 0.35 T, with the lower panel showing the zero-bias line-cut. **e**, Vertical line-cuts from **d**. **f**, Lorentzian fits (red curves) of three ZBP curves (black dots) taken from **e**, assuming a temperature broadening of 20 mK.

Recent theoretical work²⁸ has shown numerically for experimentally relevant parameters that ZBPs can also arise from local and non-topological Andreev bound states (ABS)^{16,32–35}. These local ABS appear remarkably similar in tunnelling spectroscopy to the ZBPs arising from Majorana zero-modes. In a third device, we are able to find such non-topological states by fine-tuning the gate voltages (see Extended Data Fig. 7 for specifics of all devices). Figure 5 shows the similarities and differences between ABS and Majorana ZBPs. First, Fig. 5a shows a ZBP in tunnelling spectroscopy versus B field. At a particular B field (0.7 T, red bar), the ZBP height reaches $2e^2/h$. In this device, we next vary the chemical potential by means of a voltage applied to a back-gate, producing a fairly stable (non-split) ZBP (Fig. 5b). In contrast, the ZBP is unstable against variations in tunnel-gate voltage: Fig. 5c shows that the ZBP now appears as level crossings instead of being rigidly bound to zero bias. The two different behaviours between back-gate and tunnel-gate are expected for ABSs that are localized near the tunnel barrier, as was modelled explicitly in ref. 28 (see also Extended Data Fig. 5). Liu *et al.*²⁸ show that local ABSs can have near-zero energy, which in a B field is remarkably robust against variations in chemical potential, in our experiment tuned by the back-gate. But this is only the case for the tunnel-gate voltage fine-tuned to level crossing points at zero bias. The local tunnel-gate and the global back-gate thus have distinguishably

different effects. For the Majorana case, instead of level crossing, the ZBP should remain non-split over sizable changes in tunnel-gate voltage^{14,36}, as shown in Fig. 2a and Fig. 4b.

The second fundamental difference is that the non-topological ABS ZBP height is not expected to be robustly quantized at $2e^2/h$ (refs 5, 28). Figure 5d and e shows that the ZBP height varies smoothly as a function of the back-gate voltage without any particular feature at $2e^2/h$. The ZBP height in Fig. 5a at $2e^2/h$ is just a tuned coincidence (see Extended Data Fig. 6). Note that the ZBP line-shape or temperature dependence does not discriminate between topological and non-topological cases. Both fit a Lorentzian line-shape as shown explicitly for the non-topological ABS in Fig. 5f. Thus, the temperature dependence alone cannot distinguish a Majorana origin from ABS^{7,31,32}. Only a stable quantized tunnel-conductance plateau, robust against variations in all gate voltages and magnetic field strength, can uniquely identify a topological Majorana zero-mode in tunnelling spectroscopy.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 October 2017; accepted 18 January 2018.

Published online 28 March 2018.

1. Kitaev, A. Y. Unpaired Majorana fermions in quantum wires. *Phys. Uspekhi* **44**, 131–136 (2001).
2. Mourik, V. *et al.* Signatures of Majorana fermions in hybrid superconductor-semiconductor nanowire devices. *Science* **336**, 1003–1007 (2012).
3. Law, K. T., Lee, P. A. & Ng, T. K. Majorana fermion induced resonant Andreev reflection. *Phys. Rev. Lett.* **103**, 237001 (2009).
4. Flensberg, K. Tunneling characteristics of a chain of Majorana bound states. *Phys. Rev. B* **82**, 180516 (2010).
5. Wimmer, M., Akhmerov, A. R., Dahlhaus, J. P. & Beenakker, C. W. J. Quantum point contact as a probe of a topological superconductor. *New J. Phys.* **13**, 053016 (2011).
6. Lutchyn, R. M. *et al.* Realizing Majorana zero modes in superconductor-semiconductor heterostructures. Preprint at <https://arxiv.org/abs/1707.04899> (2017).
7. Nichele, F. *et al.* Scaling of Majorana zero-bias conductance peaks. *Phys. Rev. Lett.* **119**, 136803 (2017).
8. Lutchyn, R. M., Sau, J. D. & Das Sarma, S. Majorana fermions and a topological phase transition in semiconductor-superconductor heterostructures. *Phys. Rev. Lett.* **105**, 077001 (2010).
9. Oreg, Y., Refael, G. & von Oppen, F. Helical liquids and Majorana bound states in quantum wires. *Phys. Rev. Lett.* **105**, 177002 (2010).
10. Sengupta, K., Žutić, I., Kwon, H.-J., Yakovenko, V. M. & Das Sarma, S. Midgap edge states and pairing symmetry of quasi-one-dimensional organic superconductors. *Phys. Rev. B* **63**, 144531 (2001).
11. Majorana, E. A symmetric theory of electrons and positrons. *Soryushiron Kenkyu* **63**, 149 (1981) [transl.: *Nuovo Cimento* **14**, 171–184 (1937)].
12. Read, N. & Green, D. Paired states of fermions in two dimensions with breaking of parity and time-reversal symmetries and the fractional quantum Hall effect. *Phys. Rev. B* **61**, 10267–10297 (2000).
13. Deng, M. T. *et al.* Majorana bound state in a coupled quantum-dot hybrid-nanowire system. *Science* **354**, 1557–1562 (2016).
14. Gül, Ö. *et al.* Ballistic Majorana nanowire devices. *Nature Nanotech.* **13**, 192–197 (2018).
15. Pientka, F., Kells, G., Romito, A., Brouwer, P. W. & von Oppen, F. Enhanced zero-bias Majorana peak in the differential tunneling conductance of disordered multisubband quantum-wire/superconductor junctions. *Phys. Rev. Lett.* **109**, 227006 (2012).
16. Prada, E., San-Jose, P. & Aguado, R. Transport spectroscopy of NS nanowire junctions with Majorana fermions. *Phys. Rev. B* **86**, 180503 (2012).
17. Lin, C.-H., Sau, J. D. & Das Sarma, S. Zero-bias conductance peak in Majorana wires made of semiconductor/superconductor hybrid structures. *Phys. Rev. B* **86**, 224511 (2012).
18. Rainis, D., Trifunovic, L., Klinovaja, J. & Loss, D. Towards a realistic transport modeling in a superconducting nanowire with Majorana fermions. *Phys. Rev. B* **87**, 024515 (2013).
19. Liu, C.-X., Sau, J. D. & Das Sarma, S. Role of dissipation in realistic Majorana nanowires. *Phys. Rev. B* **95**, 054502 (2017).
20. Krogstrup, P. *et al.* Epitaxy of semiconductor-superconductor nanowires. *Nat. Mater.* **14**, 400–406 (2015).
21. Gazibegovic, S. *et al.* Epitaxy of advanced nanowire quantum devices. *Nature* **548**, 434–438 (2017).
22. Gül, Ö. *et al.* Hard superconducting gap in InSb nanowires. *Nano Lett.* **17**, 2690–2696 (2017).
23. Zhang, H. *et al.* Ballistic superconductivity in semiconductor nanowires. *Nat. Commun.* **8**, 16025 (2017).

24. Takei, S., Fregoso, B. M., Hui, H.-Y., Lobos, A. M. & Das Sarma, S. Soft superconducting gap in semiconductor Majorana nanowires. *Phys. Rev. Lett.* **110**, 186803 (2013).
25. Stanescu, T. D., Tewari, S., Sau, J. D. & Das Sarma, S. To close or not to close: the fate of the superconducting gap across the topological quantum phase transition in Majorana-carrying semiconductor nanowires. *Phys. Rev. Lett.* **109**, 266402 (2012).
26. Mishmash, R. V., Aasen, D., Higginbotham, A. P. & Alicea, J. Approaching a topological phase transition in Majorana nanowires. *Phys. Rev. B* **93**, 245404 (2016).
27. Liu, C.-X., Setiawan, F., Sau, J. D. & Das Sarma, S. Phenomenology of the soft gap, zero-bias peak, and zero-mode splitting in ideal Majorana nanowires. *Phys. Rev. B* **96**, 054520 (2017).
28. Liu, C.-X., Sau, J. D., Stanescu, T. D. & Das Sarma, S. Andreev bound states versus Majorana bound states in quantum dot–nanowire–superconductor hybrid structures: trivial versus topological zero-bias conductance peaks. *Phys. Rev. B* **96**, 075161 (2017).
29. Das Sarma, S., Sau, J. D. & Stanescu, T. D. Splitting of the zero-bias conductance peak as smoking gun evidence for the existence of the Majorana mode in a superconductor–semiconductor nanowire. *Phys. Rev. B* **86**, 220506 (2012).
30. Zazunov, A., Egger, R. & Levy Yeyati, A. Low-energy theory of transport in Majorana wire junctions. *Phys. Rev. B* **94**, 014502 (2016).
31. Setiawan, F., Liu, C.-X., Sau, J. D. & Das Sarma, S. Electron temperature and tunnel coupling dependence of zero-bias and almost-zero-bias conductance peaks in Majorana nanowires. *Phys. Rev. B* **96**, 184520 (2017).
32. Kells, G., Meidan, D. & Brouwer, P. W. Near-zero-energy end states in topologically trivial spin–orbit coupled superconducting nanowires with a smooth confinement. *Phys. Rev. B* **86**, 100503 (2012).
33. Lee, E. J. H. *et al.* Spin-resolved Andreev levels and parity crossings in hybrid superconductor–semiconductor nanostructures. *Nat. Nanotech.* **9**, 79–84 (2014).
34. Pikulin, D. I., Dahlhaus, J. P., Wimmer, M., Schomerus, H. & Beenakker, C. W. J. A zero-voltage conductance peak from weak antilocalization in a Majorana nanowire. *New J. Phys.* **14**, 125011 (2012).
35. Stanescu, T. D. & Tewari, S. Disentangling Majorana fermions from topologically trivial low-energy states in semiconductor Majorana wires. *Phys. Rev. B* **87**, 140504 (2013).
36. Prada, E., Aguado, R. & San-Jose, P. Measuring Majorana nonlocality and spin structure with a quantum dot. *Phys. Rev. B* **96**, 085418 (2017).

Acknowledgements We thank M. Wimmer and Ö. Gül for discussions. This work has been supported by the European Research Council, the Dutch Organization for Scientific Research, the Office of Naval Research, the Laboratory for Physical Sciences and Microsoft Corporation Station-Q.

Author Contributions H.Z., D.X., G.W., N.v.L., J.D.S.B. and M.W.A.d.M. fabricated the devices, performed electrical measurements and analysed the experimental data. S.G., J.A.L., D.C., R.L.M.O.h.V., P.J.v.V., S.K., M.A.V., M.P., D.J.P., B.S., J.S.L., C.J.P. and E.P.A.M.B. grew the nanowires with epitaxial Al and performed the nanowire deposition. C.-X.L. and S.D.S. performed the numerical simulations. The manuscript was written by H.Z. and L.P.K. with comments from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to H.Z. (H.Zhang-3@tudelft.nl) or L.P.K. (Leo.Kouwenhoven@microsoft.com).

Reviewer Information *Nature* thanks M. Franz and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

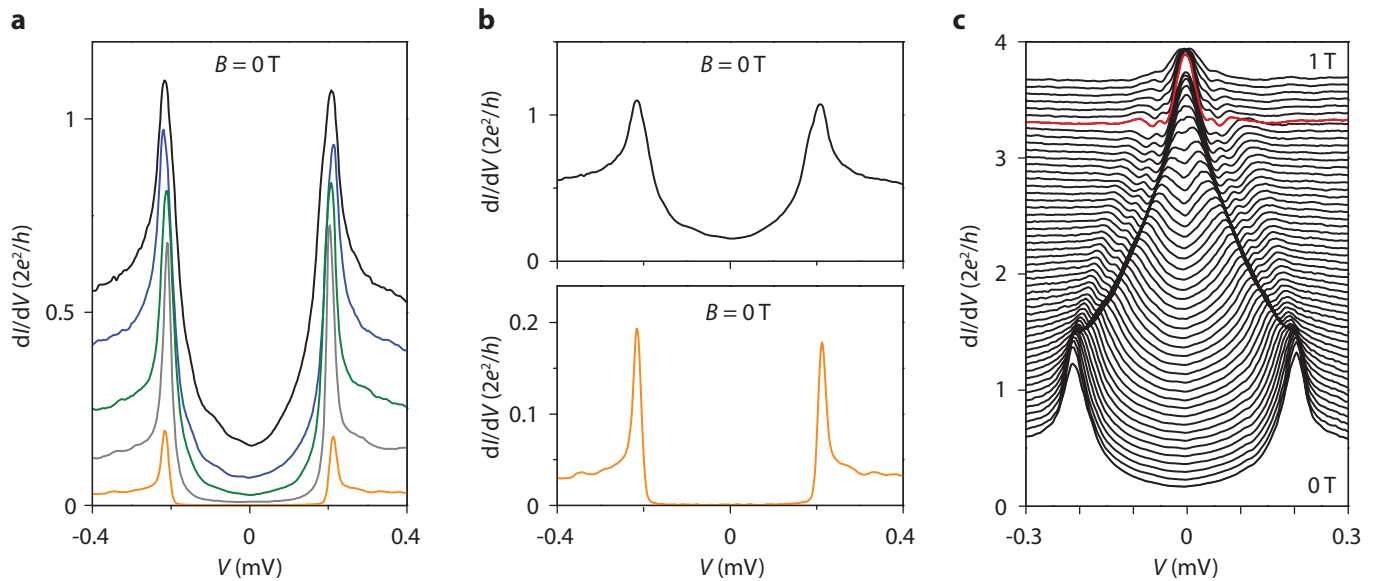
Theory model. We use the theoretical model from ref. 28 to perform numerical simulations with experimentally relevant parameters, such as the effective mass $m^* = 0.015m_e$, the spin–orbit coupling $\alpha = 0.5 \text{ eV \AA}$, the chemical potential of the normal metal lead $\mu_{\text{lead}} = 25 \text{ meV}$, the Landé g -factor $g = 40$ such that the Zeeman energy $V_Z [\text{meV}] = 1.2 \text{ B [T]}$, and the length of the nanowire $L = 1.0 \mu\text{m}$. Note that the collapse of the bulk Al superconducting gap is included explicitly in the theory to be consistent with the experimental situation in which the bulk gap collapses at about 1 T.

Lorentzian fit. We fit our ZBP line-shape with the Lorentzian formula:

$$G(V) = \frac{2e^2}{h} \frac{\Gamma^2}{\Gamma^2 + (eV)^2},$$

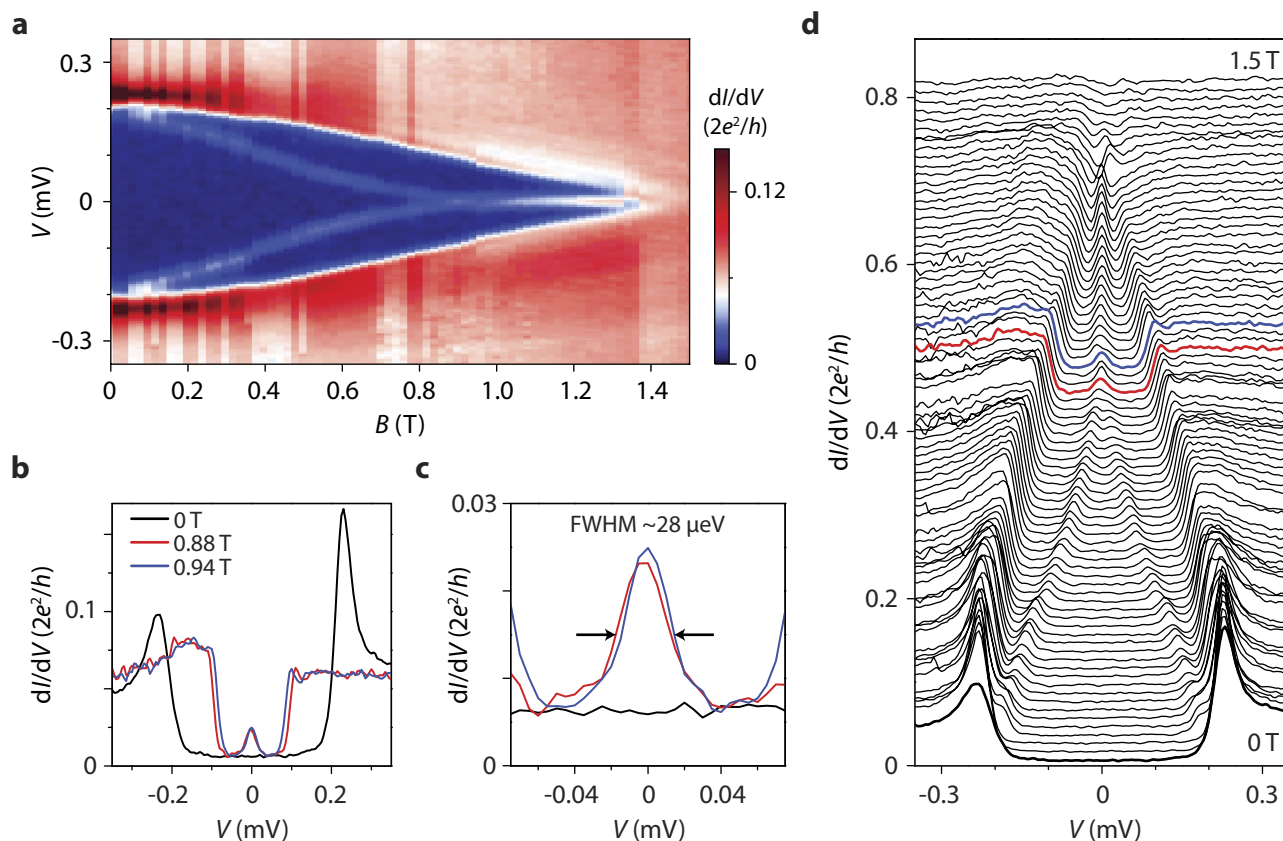
where Γ defines the tunnel coupling and FWHM of the peak, that is, 2Γ . Then we do convolution integration with the derivative of the Fermi distribution function (at 20 mK) to fit our ZBP shape. Because the FWHM of our ZBP is much larger than the thermal width, we take Γ to be roughly equal to half of the FWHM for all the fittings in Fig. 4c and Fig. 5f.

Data availability. The data that support the findings of this study are available within the paper. Additional data are available from the corresponding authors upon request.



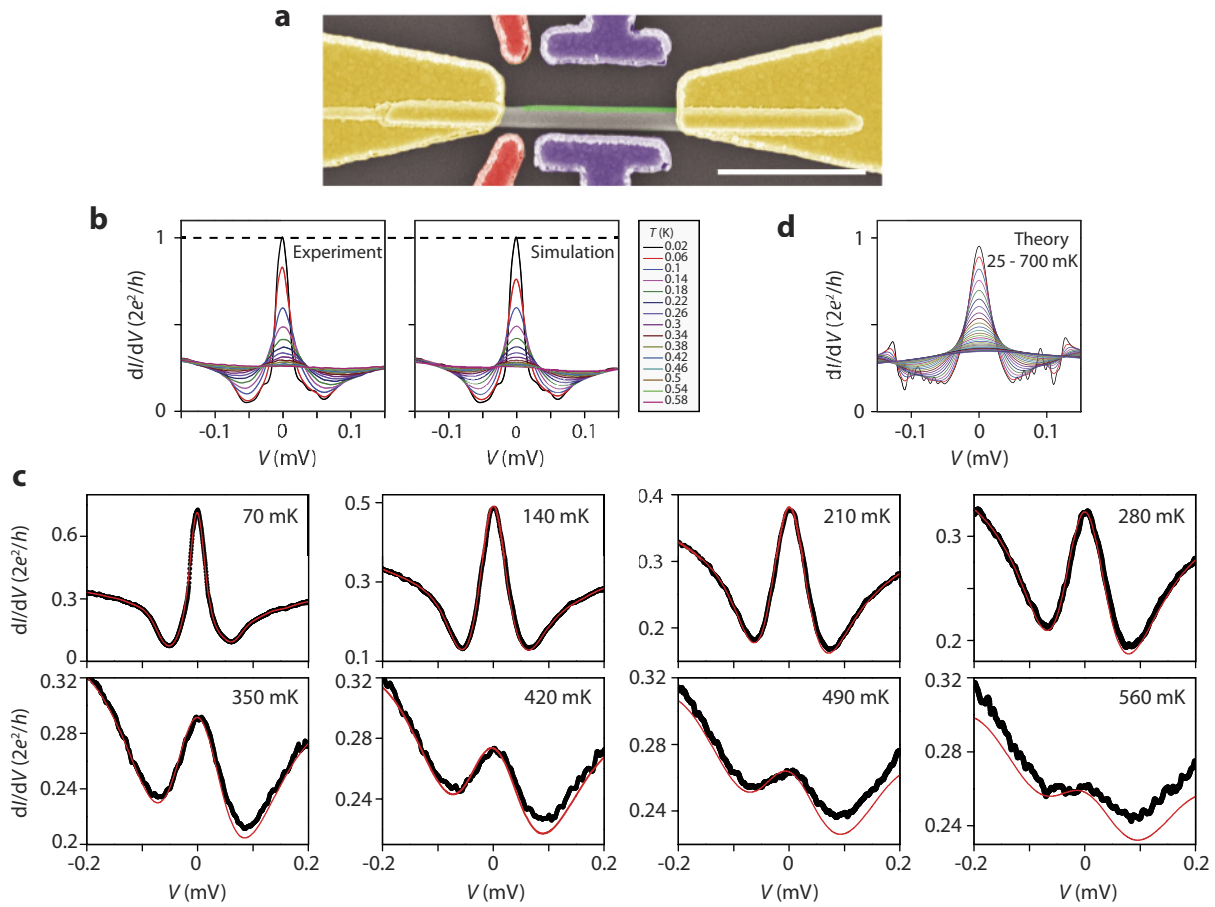
Extended Data Figure 1 | Apparent ‘soft gap’ due to large Andreev reflection. **a**, Differential conductance dI/dV of the device in Figs 1–3 (device A) as a function of bias voltage at zero magnetic field. The tunnel-gate voltage is tuned to more negative from the top curve to the bottom curve. The transmission probability of the tunnel barrier is tuned from large (black curve) to small (orange curve). In the low transmission regime (orange curve), where the above-gap conductance (about $0.03 \times 2e^2/h$) is much less than $2e^2/h$, dI/dV is proportional to the density of states in the proximitized wire part, resolving a hard superconducting gap. In the high

transmission regime (black curve), where the above-gap conductance is comparable with $2e^2/h$, the finite sub-gap conductance is due to large Andreev reflection. This ‘soft gap’ is not from dissipation, and does not affect the quantized ZBP height as shown in **c**. **b**, Re-plot of the two extreme curves from **a**, for clarity. **c**, Waterfall plot of Fig. 1b, showing all the individual curves from 0 T to 1 T in steps of 0.02 T. The curves are offset vertically by $0.066 \times 2e^2/h$ for clarity. The curve at 0 T and the red curve at 0.88 T correspond to the curves in Fig. 1c (left panel).



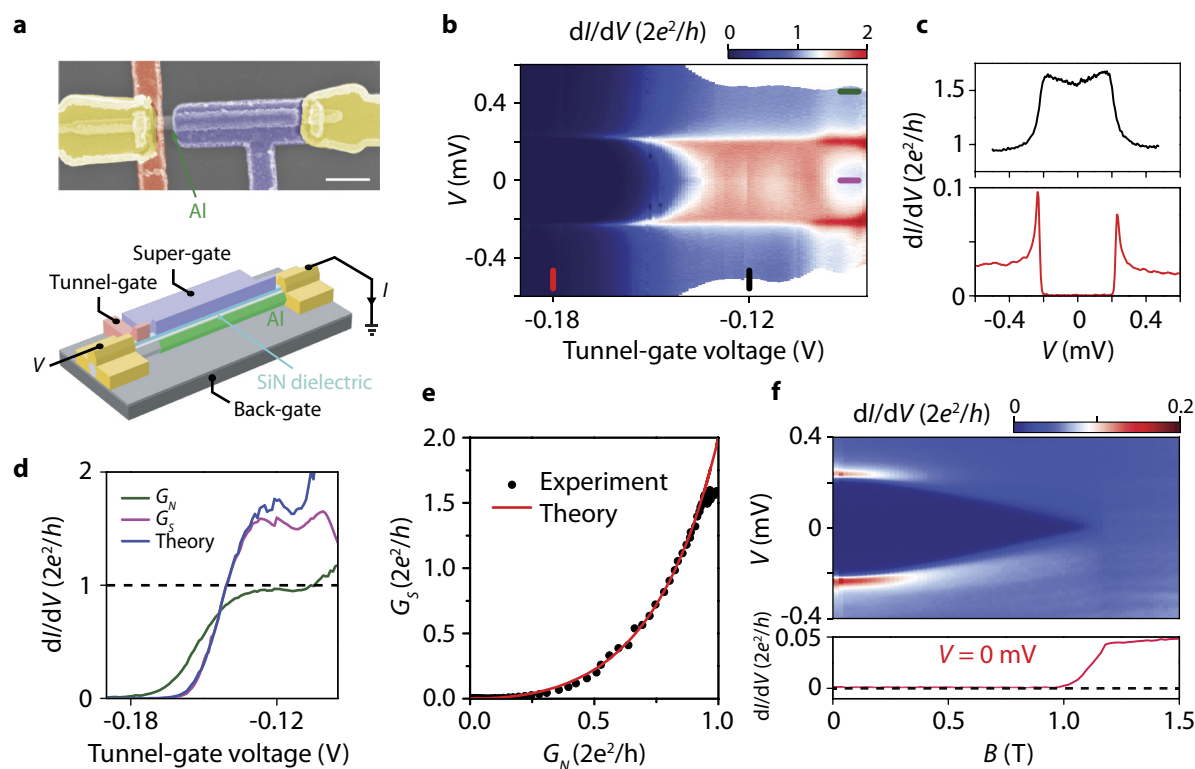
Extended Data Figure 2 | Thermal-broadened ZBP in low transmission regime. **a**, Differential conductance dI/dV of device D, as a function of B , showing a stable ZBP. **b**, Vertical line-cuts at 0 T, 0.88 T and 0.94 T. At $B = 0$ T, the above-gap conductance (approximately $0.05 \times 2e^2/h$) is much less than $2e^2/h$, which means that the device is in the low transmission regime, and thus shows a hard gap. The tiny sub-gap conductance is due to the small Andreev reflection and the noise background of the measurement equipment. The low transmission leads to a narrow ZBP width, which is negligible compared with the thermal width of $3.5k_B T$. Thus, thermal averaging suppresses the ZBP height below the quantized

value. The sub-gap conductance at finite B (for example, 0.88 T or 0.94 T), where the ZBP appears, is the same as the sub-gap conductance at zero field, indicating that the gap remains hard at high magnetic field where the Majorana state is present. **c**, The zoom-in curves show that the FWHM of the ZBP is about $28 \mu\text{eV}$, which is consistent with the combined effect of the thermal broadening ($3.5k_B T \approx 6 \mu\text{eV}$ at 20 mK), the lock-in bias voltage excitation ($5 \mu\text{eV}$) and broadening from tunnelling. This shows that the thermal broadening does indeed dominate over tunnel broadening. **d**, Waterfall plot of **a** with vertical offset of $0.01 \times 2e^2/h$ for clarity.



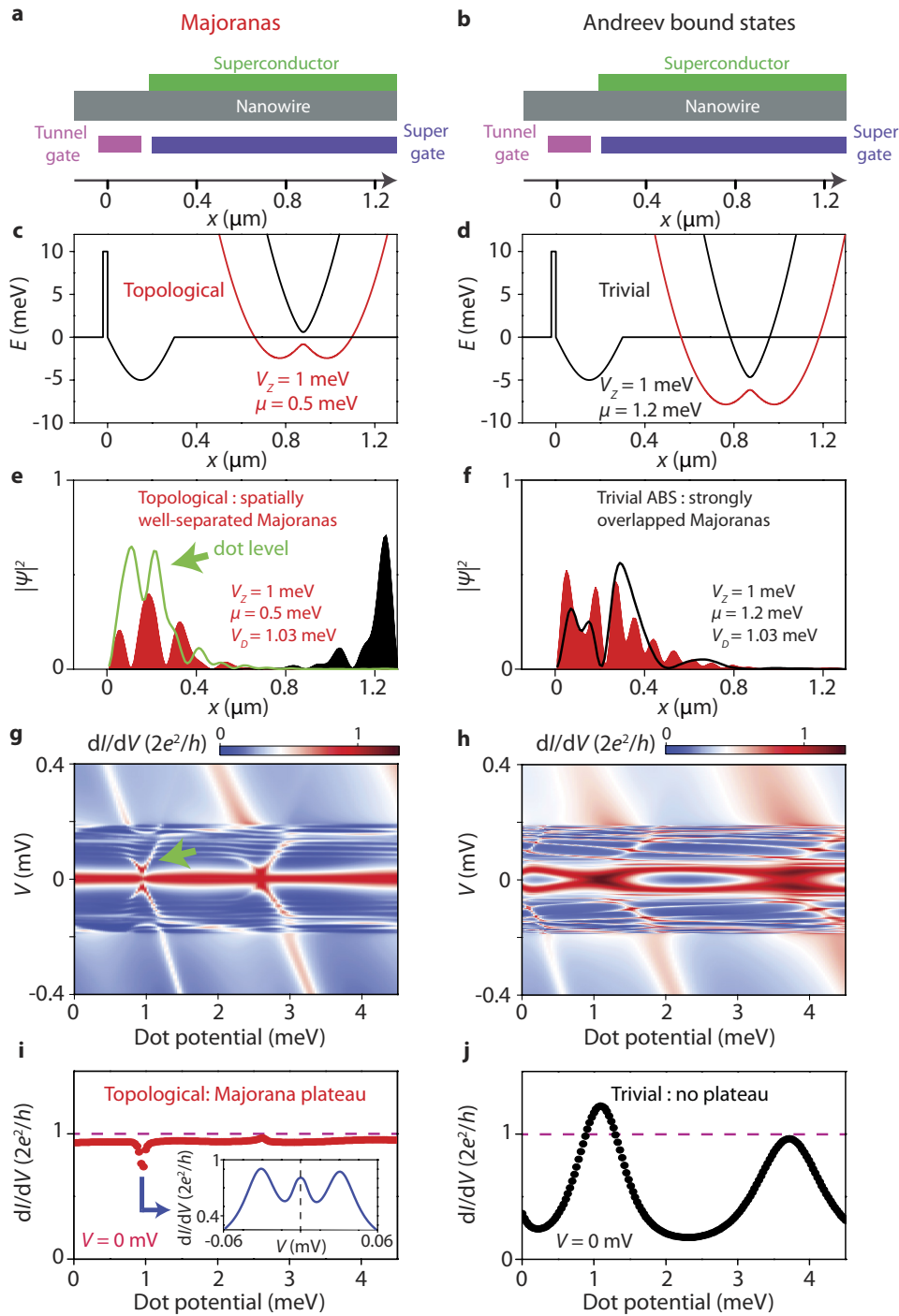
Extended Data Figure 3 | Simulation of temperature dependence on the quantized ZBP. **a**, False-colour scanning electron micrograph of device B with data shown in Fig. 4. Scale bar is 1 μm . The length of the Al section is about 0.9 μm . We calculate the dI/dV curve at high temperature by convolution of the derivative of the Fermi distribution function with the dI/dV curve at base temperature of 20 mK: $dI/dV = G(V, T) = \int_{-\infty}^{\infty} d\epsilon G(\epsilon, 0) \frac{df(\epsilon V - \epsilon, T)}{d\epsilon}$, where T is temperature, V is bias voltage, and $f(E, T)$ is the Fermi distribution function. Because we use the dI/dV curve at 20 mK as the zero-temperature data, our model only works for T sufficiently

larger than 20 mK, that is, $T > 50$ mK. **b**, Comparison between the experimental data (left, taken from Fig. 4d) and theory simulations, for different temperatures. **c**, Several typical curves at different temperatures; black traces are the experimental data, and the red curves are the theory simulations with no fitting parameters. The agreement between simulation and experiment indicates that thermal averaging effect is the dominating effect that smears out the ZBP at high temperature. **d**, Temperature dependence of the ZBP taken from our theory model: Fig. 1c (right panel). The temperature varies from 25 mK to 700 mK in steps of 23 mK.



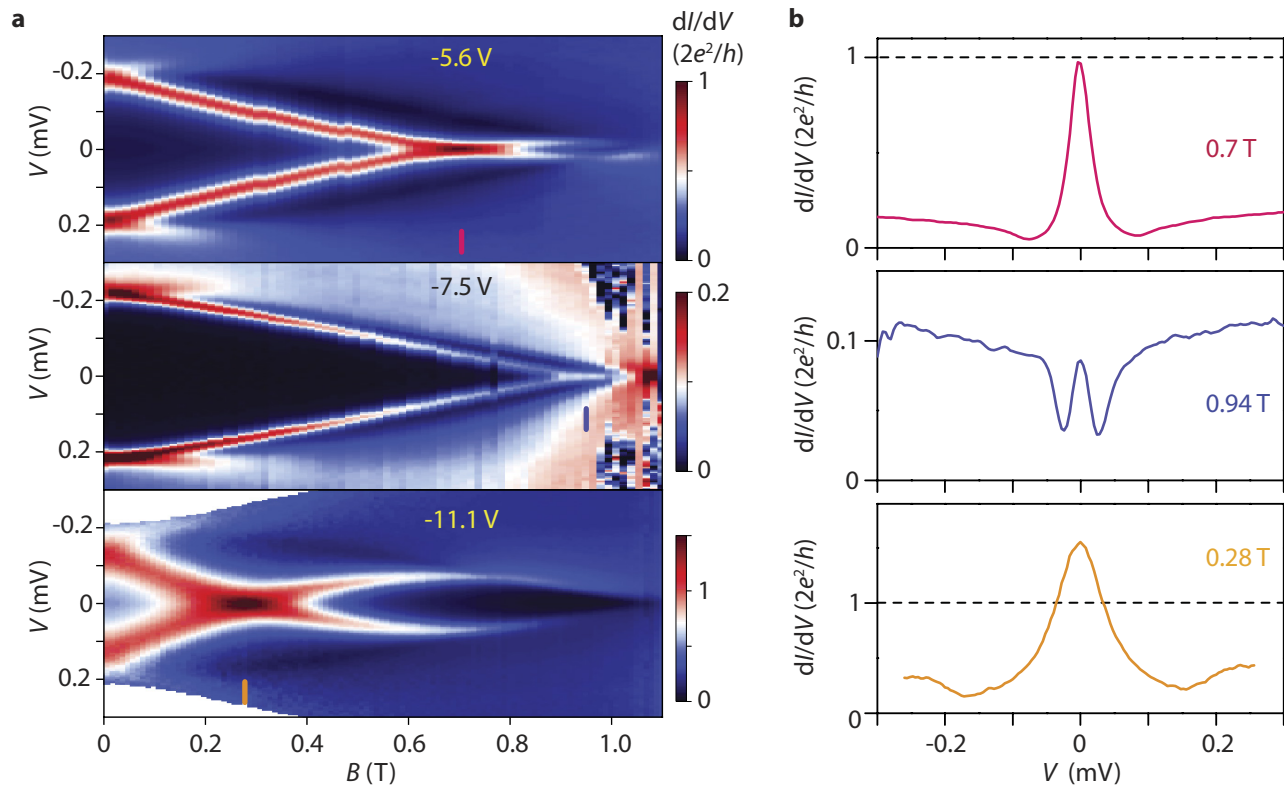
Extended Data Figure 4 | Perfect ballistic Andreev transport in InSb-Al nanowires. **a**, False-colour scanning electron micrograph of the device in Fig. 5 (device C). Scale bar is 500 nm. Electrical contacts and top gates are Cr/Au. Lower panel shows the device schematic and measurement set-up. The two top-gates (tunnel-gate and super-gate) are separated from the nanowire by 30-nm-thick SiN dielectric. The global back gate is p-doped Si covered by 285-nm-thick SiO₂ dielectric. **b**, Differential conductance dI/dV , as a function of bias voltage (V) and tunnel-gate voltage at zero field. No localization effect (conductance resonances or quantum-dot-induced Coulomb blockade) is observed. **c**, Vertical line-cuts from **b** at tunnel-gate voltage of -0.18 V (lower panel) and -0.12 V (upper panel), showing a hard superconducting gap in the low transmission regime (lower panel) and strong Andreev enhancement in the open regime

(upper panel). **d**, Horizontal line-cuts from **c** for $V = 0$ mV (pink, sub-gap conductance, G_N) and $V = 0.45$ mV (green, above-gap conductance, G_N). The blue curve is the calculated sub-gap conductance using $G_S = 4e^2/h \times T^2/(2 - T)^2$, where the transmission T is extracted from the above-gap conductance: $G_N = (2e^2/h) \times T$. **e**, Sub-gap conductance G_S as a function of G_N (black dots) and the theory prediction (red curve): $G_S = 2G_N^2/(2 - G_N)^2$, with G_S and G_N in unit of $2e^2/h$. Both **d** and **e** show perfect agreement between theory and experiment. This indicates that the sub-gap conductance is indeed dominated by the Andreev reflection, that is, without contributions from sub-gap states. **f**, Magnetic field dependence of the hard gap. Lower panel shows the zero-bias line-cut. The gap remains hard up to 1 T, where the bulk superconducting gap closes.



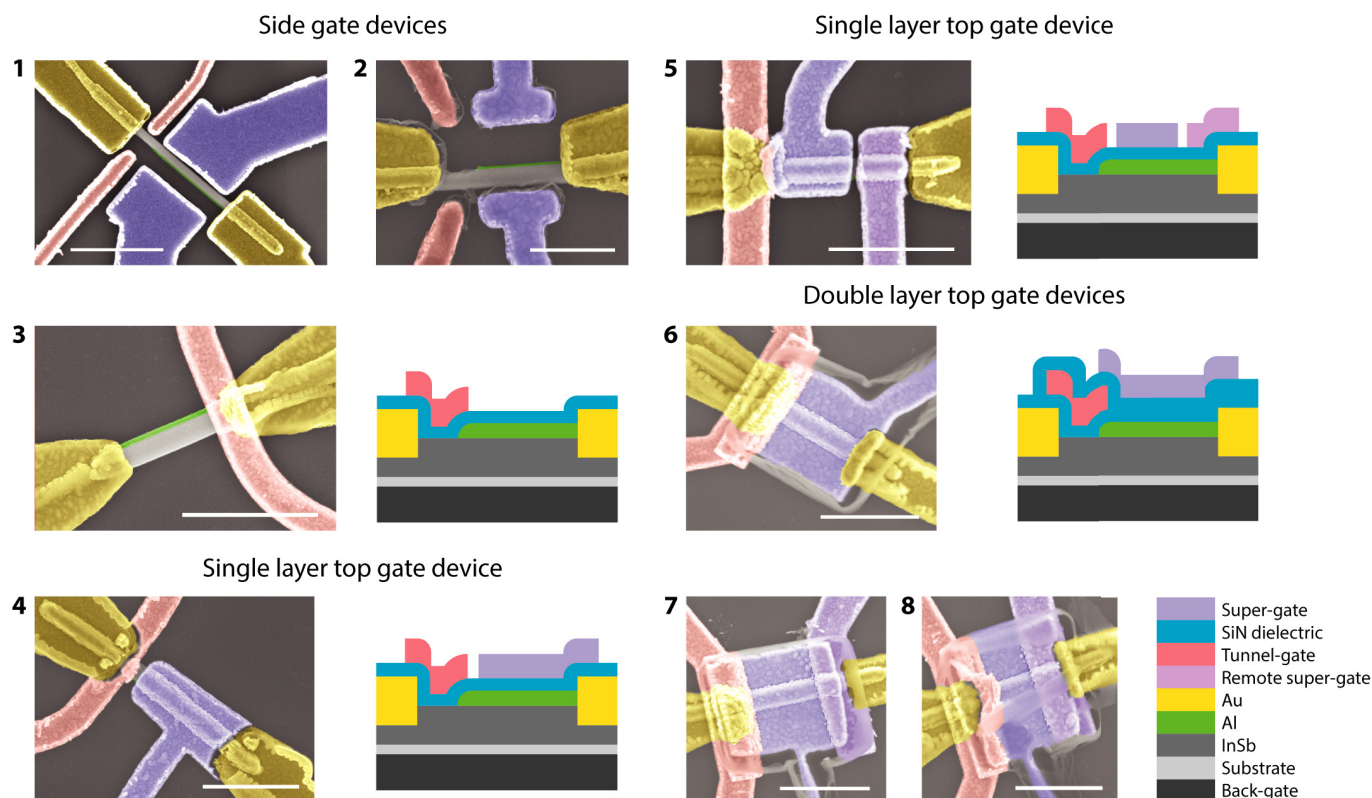
Extended Data Figure 5 | Majoranas versus trivial Andreev bound states. **a, b**, Schematics of a Majorana nanowire device. The only difference between the left column (Majorana) and right column (ABS) is the chemical potential, as shown in **c** and **d**. **c, d**, Potential profile in the device. The tunnel barrier height is 10 meV and the width is 10 nm. The dot potential shape is $E(x) = -V_D \sin(\pi x/l_{\text{dot}})$, for x between 0 and $0.3 \mu\text{m}$, where the length of the dot (l_{dot}) is $0.3 \mu\text{m}$, and V_D is the dot depth which can be tuned by the nearby gate, that is, the tunnel-gate. The rest of the flat nanowire segment is $1 \mu\text{m}$ long. We assume a pairing potential $\Delta = 0.2 \text{ meV}$, with a spin-orbit coupling of 0.5 eV \AA . We set the Zeeman energy to be 1 meV, so the chemical potential of 0.5 meV (left) corresponds to the topological regime, and 1.2 meV (right) corresponds to the trivial regime, based on the topological condition $V_Z > \sqrt{\mu^2 + \Delta^2}$, where μ is chemical potential. **e, f**, Spatial distribution of the Majorana and ABS wavefunctions in the topological and trivial regime. In the topological regime, two spatially well separated Majoranas (red and black) are localized at the two ends of the topological section. In the trivial regime,

the Andreev bound state, which can be considered as two strongly overlapped Majoranas (red and black), is localized near the tunnel barrier. **g, h**, The Majorana ZBP remains non-split against the change of dot potential, regardless of the energy of the dot level. The green arrow indicates one bound state in the dot, whose wavefunction $|\psi^2|$ is shown in **e** (green curve). When this dot level moves down, it is repelled from zero energy, where the Majorana ZBP remains bound to zero (inset of **i**). On the contrary, the ABS-induced ZBP is not robust at all and only shows up at the crossing points of two Andreev levels. This is because the tunnel-gate tunes the dot potential, which therefore affects the energy of the localized ABS near the tunnel barrier. **i, j**, The Majorana ZBP height shows a quantized plateau at $2e^2/h$ by tuning the dot potential with tunnel-gate. The ZBP height drops from the quantized value (inset) when the ABS-dot level moves towards zero, which effectively squeezes the ZBP-width such that the thermal averaging effect starts to dominate. The ABS zero-bias conductance does not show a plateau, but instead varies between 0 and $4e^2/h$.



Extended Data Figure 6 | Magnetic field dependence of trivial Andreev bound states. **a**, Top panel is a re-plot of the trivial ABS data in Fig. 5a. Middle and bottom panels are the ZBP data at different back-gate voltages

(labelled in the panels). **b**, Line-cuts of the ZBP data from **a**. The ZBP height varies with back-gate voltages and can exceed $2e^2/h$. The ZBP height at $2e^2/h$ here is just a tuned coincidence.



Extended Data Figure 7 | Specifics of devices. We fabricated and tested many (over 60) devices out of which we selected 11 devices that showed good basic transport with all gates being fully functional. These were used for extensive measurements. Although most of these devices show ZBPs after tuning gate voltages and magnetic field, only two devices (presented in the main text: Figs 1–3 for device A and Fig. 4 for device B) show a quantized ZBP plateau. All other devices show trivial ZBPs similar to Fig. 5 (from device C). Scanning electron microscope images of devices A, B and C are shown in Fig. 1a, Extended Data Fig. 3a and Extended Data Fig. 4a, respectively. Here we show the scanning electron microscope images of the other eight devices, which we have explored extensively, but without finding a quantized ZBP plateau. Devices 1 and 2 are side-gate

devices. Device 3 has a top tunnel-gate separated from the nanowire by 30-nm-thick SiN dielectric, and a global back-gate separated by 285-nm-thick SiO₂. Devices 4 and 5 have tunnel-gate and super-gate on top separated from the nanowire by 30-nm-thick SiN dielectric. Devices 6 to 8 have two layers of top-gate. The bottom layer has a tunnel-gate separated by 30-nm-thick SiN dielectric while the top layer has super-gates separated by 30-nm-thick SiN from the bottom layer. The scale bar is 1 μm for all devices, except for device 2, which is 500 nm. It would be informative to perform Schrodinger–Poisson calculations on these different device geometries to determine the self-consistent potential landscape and find out which geometry suppresses a local potential dip near the tunnel barrier.

Correlated insulator behaviour at half-filling in magic-angle graphene superlattices

Yuan Cao¹, Valla Fatemi¹, Ahmet Demir¹, Shiang Fang², Spencer L. Tomarken¹, Jason Y. Luo¹, Javier D. Sanchez-Yamagishi², Kenji Watanabe³, Takashi Taniguchi³, Efthimios Kaxiras^{2,4}, Ray C. Ashoori¹ & Pablo Jarillo-Herrero¹

A van der Waals heterostructure is a type of metamaterial that consists of vertically stacked two-dimensional building blocks held together by the van der Waals forces between the layers. This design means that the properties of van der Waals heterostructures can be engineered precisely, even more so than those of two-dimensional materials¹. One such property is the ‘twist’ angle between different layers in the heterostructure. This angle has a crucial role in the electronic properties of van der Waals heterostructures, but does not have a direct analogue in other types of heterostructure, such as semiconductors grown using molecular beam epitaxy. For small twist angles, the moiré pattern that is produced by the lattice misorientation between the two-dimensional layers creates long-range modulation of the stacking order. So far, studies of the effects of the twist angle in van der Waals heterostructures have concentrated mostly on heterostructures consisting of monolayer graphene on top of hexagonal boron nitride, which exhibit relatively weak interlayer interaction owing to the large bandgap in hexagonal boron nitride^{2–5}. Here we study a heterostructure consisting of bilayer graphene, in which the two graphene layers are twisted relative to each other by a certain angle. We show experimentally that, as predicted theoretically⁶, when this angle is close to the ‘magic’ angle the electronic band structure near zero Fermi energy becomes flat, owing to strong interlayer coupling. These flat bands exhibit insulating states at half-filling, which are not expected in the absence of correlations between electrons. We show that these correlated states at half-filling are consistent with Mott-like insulator states, which can arise from electrons being localized in the superlattice that is induced by the moiré pattern. These properties of magic-angle-twisted bilayer graphene heterostructures suggest that these materials could be used to study other exotic many-body quantum phases in two dimensions in the absence of a magnetic field. The accessibility of the flat bands through electrical tunability and the bandwidth tunability through the twist angle could pave the way towards more exotic correlated systems, such as unconventional superconductors and quantum spin liquids.

Exotic quantum phenomena, such as superconductivity and the fractional quantum Hall effect, often occur in condensed-matter systems and other systems with a high density of states. One way of creating a high density of states is to have ‘flat’ bands, which have weak dispersion in momentum space, with the kinetic energy of the electron set by the bandwidth W . When the Fermi level lies within the flat bands, Coulomb interactions (U) can greatly exceed the kinetic energy of the electrons and drive the system into various strongly correlated phases ($U/W \gg 1$)^{7–11}. The study of such flat-band systems in bulk materials is therefore scientifically important, and the search for new flat-band systems, such as in kagome and Lieb lattices and in heavy-fermion systems, is ongoing^{7–12}.

Recent advances in two-dimensional materials have provided a new route to achieving flat bands. An inherent advantage of two-dimensional materials is that the chemical potential of electrons can be tuned continuously via the electric-field effect without introducing extra disorder. In a twisted van der Waals heterostructure, the mismatch between two similar lattices generates a moiré pattern (Fig. 1b). This additional periodicity, which can have a length scale orders of magnitude larger than that of the underlying atomic lattices, has been shown to create a fractal energy spectrum in a strong magnetic field^{2–4}. In twisted layers, the interlayer hybridization is modulated by the moiré pattern as well. As an example, the band structure of twisted bilayer graphene (TBG) can be tailored to generate bandgaps and band curvatures that are otherwise absent^{6,13–17}. Although the well-known building blocks for van der Waals heterostructures, such as graphene and transition-metal dichalcogenides, do not have intrinsic flat bands at low energies, it has been predicted theoretically that flat bands could exist in TBG^{6,14–16,18}. Here we demonstrate experimentally that when the twist angle of TBG is close to the theoretically predicted magic angle, the interlayer hybridization induces nearly flat low-energy bands. This quenching of the quantum kinetic energy leads to a correlated insulating phase at half-filling of these flat bands, which is indicative of a Mott-like insulator in the localized flat bands.

To zeroth order, the low-energy band structure of TBG can be considered as two sets of monolayer-graphene Dirac cones rotated about the Γ point in the Brillouin zone by the twist angle θ (Fig. 1d)⁶. The difference between the two K (or K') wavevectors gives rise to the mini Brillouin zone (shown as a small hexagon), which is generated from the reciprocal lattice of the moiré superlattice (Fig. 1d). The Dirac cones near either the K or K' valley mix through interlayer hybridization, whereas interactions between distant Dirac cones are suppressed exponentially^{6,13}. As a result, the valley remains (for all practical purposes) a good quantum number. Two experimentally verified consequences of this hybridization are the energy gaps that open near the intersection of the Dirac cones and the renormalization of the Fermi velocity

$$v_F = \frac{1}{\hbar} |\nabla_k E_k|_{k=K,K'}$$

at the Dirac points^{13,19–21}.

The theoretically calculated magic angles $\theta_{\text{magic}}^{(i)}$, with $i = 1, 2, \dots$, are a series of twist angles at which the Fermi velocity at the Dirac points becomes zero⁶. The resulting low-energy bands near these twist angles are confined to less than about 10 meV. The flattening of the energy bands near the magic angle can be understood qualitatively from the competition between the kinetic energy and the interlayer hybridization energy (Fig. 1e–g). Intuitively, when the hybridization energy $2w$ is comparable to or larger than $\hbar v_0 k_\theta$, where $v_0 = 10^6 \text{ m s}^{-1}$ is the Fermi velocity of graphene, $k_\theta \approx G_K \theta$ is the momentum displacement of the

¹Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. ³National Institute for Materials Science, Namiki 1-1, Tsukuba, Ibaraki 305-0044, Japan. ⁴John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

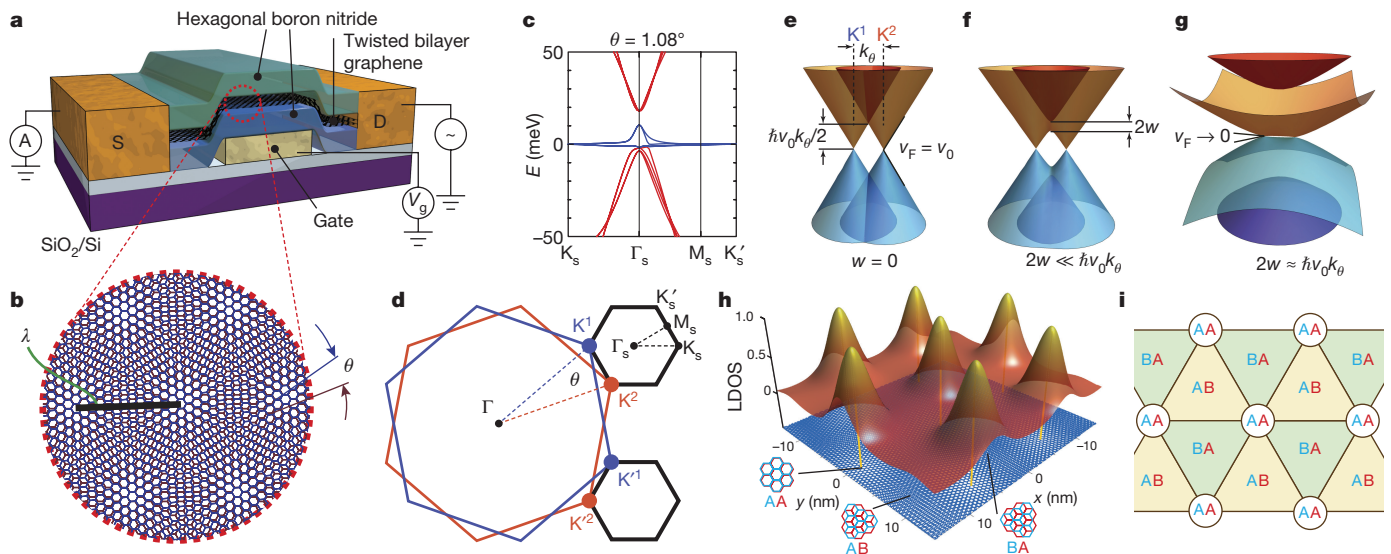


Figure 1 | Electronic band structure of twisted bilayer graphene (TBG). **a**, Schematic of the TBG devices. The TBG is encapsulated in hexagonal boron nitride flakes with thicknesses of about 10–30 nm. The devices are fabricated on SiO₂/Si substrates. The conductance is measured with a voltage bias of 100 μ V while varying the local bottom gate voltage V_g . ‘S’ and ‘D’ are the source and drain contacts, respectively. **b**, The moiré pattern as seen in TBG. The moiré wavelength is $\lambda = a/[2\sin(\theta/2)]$, where $a = 0.246$ nm is the lattice constant of graphene and θ is the twist angle. **c**, The band energy E of magic-angle ($\theta = 1.08^\circ$) TBG calculated using an *ab initio* tight-binding method. The bands shown in blue are the flat bands that we study. **d**, The mini Brillouin zone is constructed from the difference between the two K (or K’) wavevectors for the two layers.

Dirac cones, $G_K = 4\pi/(3\alpha)$ is the magnitude of the wavevector Γ –K of graphene, $\alpha = 0.246$ nm is the lattice constant of graphene and $\hbar = h/(2\pi)$ is the reduced Planck constant, the lower of the hybridized states is pushed to and crosses zero energy. A mathematical derivation of the magic-angle condition⁶ gives the first magic angle, $\theta_{\text{magic}}^{(1)} = \sqrt{3}w/(\hbar v_0 G_K) \approx 1.1^\circ$. In Fig. 1c we show an *ab initio* tight-binding calculation¹⁶ of the band structure for $\theta = 1.08^\circ$. The flat bands (coloured blue) have a bandwidth of 12 meV for the $E > 0$ branch and 2 meV for the $E < 0$ branch (where E is the band energy). From a band-theory point of view, the flat bands should have localized wavefunction profiles in real space. In Fig. 1h we show the local density of states calculated for the flat bands. The wavefunctions are indeed highly concentrated in the regions with AA stacking, whereas small but non-zero amplitudes on the AB and BA regions connect the AA regions and endow the bands with weak dispersion^{6,15,18}. A brief discussion about the topological structure of the bands near the first magic angle is given in Methods and Extended Data Fig. 1.

For the experiment, we fabricated high-quality encapsulated TBG devices with the twist angle controlled to an accuracy of about 0.1° – 0.2° using a previously developed ‘tear and stack’ technique^{13,17,22}. We measured four devices with twist angles near the first magic angle $\theta_{\text{magic}}^{(1)} \approx 1.1^\circ$. In Fig. 2a we show the low-temperature two-probe conductance of device D1 as a function of carrier density n . For $n \approx \pm n_s = \pm 2.7 \times 10^{12} \text{ cm}^{-2}$ (four electrons per moiré unit cell for $\theta = 1.08^\circ$), the conductance is zero over a wide range of densities. Here, n_s refers to the density that is required to fill the mini Brillouin zone, accounting for spin and valley degeneracies (see Methods). These insulating states have been explained previously as hybridization-induced bandgaps above and below the lowest-energy superlattice bands, and are hereafter referred to as ‘superlattice gaps’¹³. The thermal activation gaps are measured to be about 40 meV (see Methods)^{13,17}. The twist angle can be estimated from the density that is required to reach the superlattice gaps, which we find to be $\theta = 1.1^\circ \pm 0.1^\circ$ for all of the devices reported here.

Hybridization occurs between Dirac cones within each valley, whereas intervalley processes are strongly suppressed. K_s, K'_s, M_s and Γ_s denote points in the mini Brillouin zone. **e–g**, Illustration of the effect of interlayer hybridization for $w = 0$ (**e**), $2w \ll \hbar v_0 k_\theta$ (**f**) and $2w \approx \hbar v_0 k_\theta$ (**g**); $v_0 = 10^6 \text{ m s}^{-1}$ is the Fermi velocity of graphene. **h**, Normalized local density of states (LDOS) calculated for the flat bands with $E > 0$ at $\theta = 1.08^\circ$. The electron density is strongly concentrated at the regions with AA stacking order, whereas it is mostly depleted at AB- and BA-stacked regions. See Extended Data Fig. 6 for the density of states versus energy at the same twist angle. **i**, Top view of a simplified model of the stacking order.

Another pair of insulating states occurs for a narrower density range, near half the superlattice density: $n \approx \pm n_s/2 = \pm 1.4 \times 10^{12} \text{ cm}^{-2}$ (two electrons per moiré unit cell). These insulating states have a much smaller energy scale. This behaviour is markedly different from all other zero-field insulating behaviours reported previously, which occur at integer multiples of $\pm n_s$ (refs 13, 17). We refer to the states that occur near $\pm n_s/2$ as ‘half-filling insulating states’. They are observed at roughly the same density for all four devices (Fig. 2a, inset). In Fig. 2b–d we show the conductance of the half-filling states in device D1 at different temperatures. Above 4 K, the system behaves as a metal, exhibiting decreasing conductance with increasing temperature. A metal–insulator transition occurs at around 4 K. The conductance drops substantially from 4 K to 0.3 K, with the minimum value decreasing by 1.5 orders of magnitude. An Arrhenius fit yields a thermal activation gap of about 0.3 meV for the half-filling states, two orders of magnitude smaller than those of the superlattice gaps. At the lowest temperatures, the system can be limited by conduction through charge puddles, resulting in deviation from the Arrhenius fit.

To confirm the existence of the half-filling states, we performed capacitance measurements on device D2 using an a.c. low-temperature capacitance bridge (Extended Data Fig. 2)²³. The real and imaginary components of the a.c. measurement provide information about the change in capacitance and the loss tangent of the device, respectively. The latter signal is tied to the dissipation in the device due to its resistance²³. Device D2 exhibits a reduction in capacitance and strong enhancement of dissipation at $\pm n_s/2$ (Fig. 3a), in agreement with an insulating phase that results from the suppression of the density of states. The insulating state at $-n_s/2$ is weaker and visible only in the dissipation data. The observation of capacitance reduction (that is, suppression of density of states) for only the n-side half-filling state in this device may be due to an asymmetric band structure or the quality of the device. The reduction (enhancement) in capacitance (dissipation) vanishes when the device is warmed up from 0.3 K to about 2 K, consistent with the behaviour observed in transport measurements.

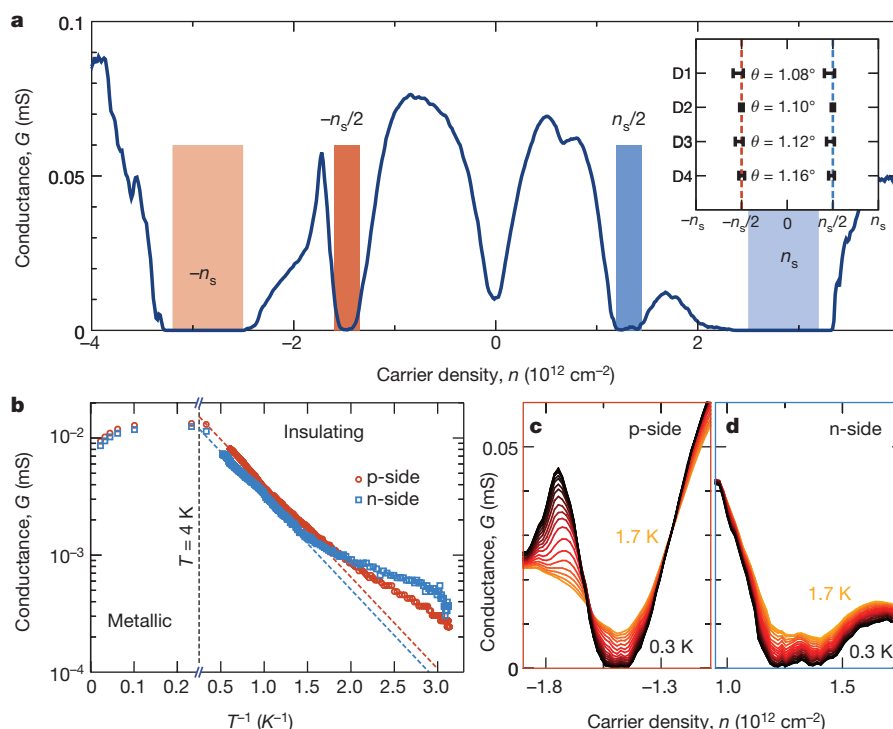


Figure 2 | Half-filling insulating states in magic-angle TBG. **a**, Measured conductance G of magic-angle TBG device D1 with $\theta = 1.08^\circ$ and $T = 0.3$ K. The Dirac point is located at $n = 0$. The lighter-shaded regions are superlattice gaps at carrier density $n = \pm n_s = \pm 2.7 \times 10^{12} \text{ cm}^{-2}$. The darker-shaded regions denote half-filling states at $\pm n_s/2$. The inset shows the density locations of half-filling states in the four different devices.

The emergence of half-filling states is not expected in the absence of interactions between electrons and appears to be correlated with the narrow bandwidth near the first magic angle. In our experiment, several separate pieces of evidence support the presence of flat bands. First, we measured the temperature dependence of the amplitude of Shubnikov–de Haas oscillations in device D1, from which we extracted the effective mass of the electron, m^* (Fig. 3b; see Methods and Extended Data Fig. 3 for analysis). For a Dirac spectrum with eight-fold degeneracy (spin, valley and layer), we expect that $m^* = \sqrt{\hbar^2 n / (8\pi v_F^2)}$, which scales as $1/v_F$. The large measured m^* near charge neutrality in device D1 indicates a reduction in v_F by a factor of 25 compared to monolayer graphene ($4 \times 10^4 \text{ m s}^{-1}$ compared to 10^6 m s^{-1}). This large reduction in the Fermi velocity is a characteristic that is expected for flat bands. Second, we analysed the capacitance data of device D2 near the Dirac point (Fig. 3a) and found that v_F needs to be reduced to about $0.15v_0$ for a good fit to the data (Methods, Extended Data Fig. 1b). Third, another direct manifestation of flat bands is the flattening of the conductance minimum at charge neutrality above a temperature of 40 K (thermal energy $kT = 3.5 \text{ meV}$), as seen in Fig. 3c. Although the conductance minimum in monolayer graphene can be observed clearly even near room temperature, it is smeared out in magic-angle TBG when the thermal energy kT becomes comparable to $v_F \hbar / 2 \approx 4 \text{ meV}$ —the energy scale that spans the Dirac-like portion of the band (Fig. 1c)^{24–26}.

Owing to the localized nature of the electrons, a plausible explanation for the gapped behaviour at half-filling is the formation of a Mott-like insulator driven by Coulomb interactions between electrons^{27,28}. To this end, we consider a Hubbard model on a triangular lattice, with each site corresponding to a localized region with AA stacking in the moiré pattern (Fig. 1i). In Fig. 3d we show the bandwidth of the $E > 0$ branch of the low-energy bands for $0.04^\circ < \theta < 2^\circ$ that we calculated numerically using a continuum model of TBG⁶. The bandwidth W is strongly suppressed near the magic angles. The on-site Coulomb energy U of each site is estimated to be $e^2 / (4\pi\epsilon d)$, where d is the effective linear

See Methods for a definition of the error bars. **b**, Minimum conductance values in the p-side (red) and n-side (blue) half-filling states in device D1. The dashed lines are fits of $\exp[-\Delta/(2kT)]$ to the data, where $\Delta \approx 0.31 \text{ meV}$ is the thermal activation gap. **c**, **d**, Temperature-dependent conductance of D1 for temperatures from about 0.3 K (black) to 1.7 K (orange) near the p-side (**c**) and n-side (**d**) half-filling states.

dimension of each site (with the same length scale as the moiré period), ϵ is the effective dielectric constant including screening and e is the electron charge. Combining ϵ and the dependence of d on twist angle into a single constant κ , we write $U = e^2 \theta / (4\pi\epsilon_0 \kappa a)$, where $a = 0.246 \text{ nm}$ is the lattice constant of monolayer graphene. In Fig. 3d we plot the on-site energy U versus θ for $\kappa = 4–20$. As a reference, $\kappa = 4$ if we assume $\epsilon = 10\epsilon_0$ and d is 40% of the moiré wavelength. For a range of possible values of κ it is therefore reasonable that $U/W > 1$ occurs near the magic angles and results in half-filling Mott-like gaps²⁷. However, the realistic scenario is much more complicated than these simplistic estimates; a complete understanding requires detailed theoretical analyses of the interactions responsible for the correlated gaps.

The Shubnikov–de Haas oscillation frequency f_{SDH} (Fig. 3b) also supports the existence of Mott-like correlated gaps at half-filling. Near the charge neutrality point, the oscillation frequency closely follows $f_{\text{SDH}} = \phi_0 |n| / M$ where $\phi_0 = h/e$ is the flux quantum and $M = 4$ indicates the spin and valley degeneracies. However, at $|n| > n_s/2$, we observe oscillation frequencies that corresponds to straight lines, $f_{\text{SDH}} = \phi_0 (|n| - n_s/2) / M$, in which M has a reduced value of 2. Moreover, these lines extrapolate to zero exactly at the densities of the half-filling states, $n = \pm n_s/2$. These oscillations point to small Fermi pockets that result from doping the half-filling states, which might originate from charged quasiparticles near a Mott-like insulator phase²⁹. The halved degeneracy of the Fermi pockets might be related to the spin–charge separation that is predicted in a Mott insulator²⁹. These results are also supported by Hall measurements at 0.3 K (Extended Data Fig. 4; see Methods for discussion), which show a ‘resetting’ of the Hall densities when the system is electrostatically doped beyond the Mott-like states.

The half-filling states at $\pm n_s/2$ are suppressed by the application of a magnetic field. In Fig. 4a, b we show that both insulating phases start to conduct at a perpendicular field of $B = 4 \text{ T}$ and recover normal conductance by $B = 8 \text{ T}$. A similar effect is observed for an in-plane magnetic field (Extended Data Fig. 5d). The insensitivity to field

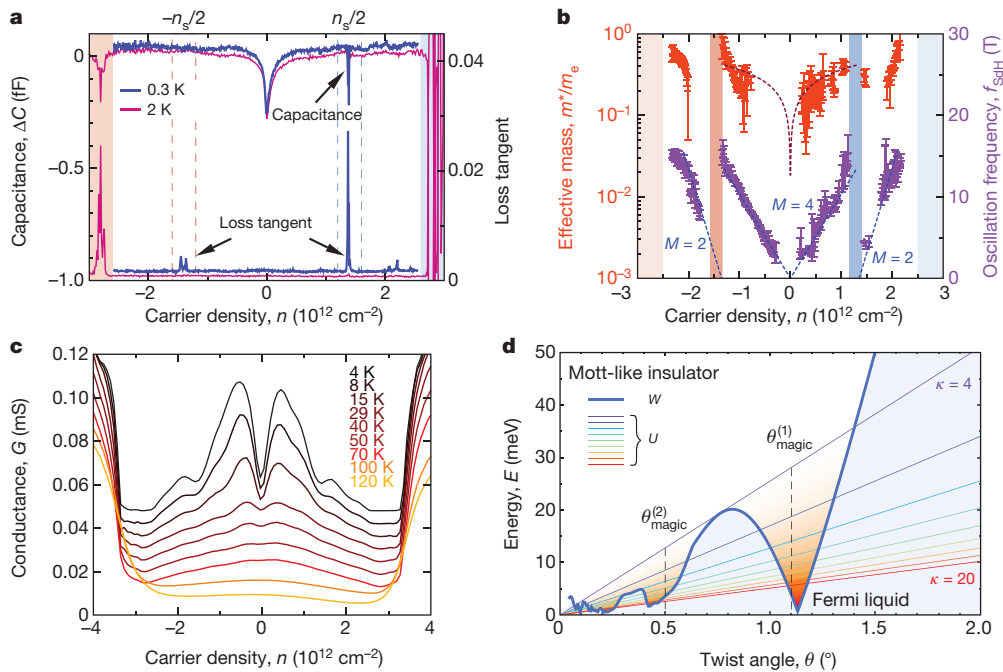


Figure 3 | Flat bands in magic-angle TBG. **a**, Capacitance measurements of device D2 at 0.3 K (blue) and 2 K (pink). The change in the measured capacitance (ΔC ; upper traces) is plotted on the left axis and the loss tangent (lower traces) is shown on the right axis. For densities corresponding to half-filling ($\pm n_s/2$; dashed vertical lines), a reduction in ΔC (on the p-side only) and an enhancement in loss tangent (on both sides) are observed in the 0.3 K data. These effects disappear in the 2 K measurements. **b**, The effective mass m^* and oscillation frequency f_{SDH} as extracted from temperature-dependent Shubnikov–de Haas oscillations. The fitting curve (red dashed line) is $m^* = \sqrt{\hbar^2 |n| / (8\pi v_F^2)}$, assuming a uniform Fermi velocity of v_F . For magic-angle device D1, the estimated Fermi velocity of $v_F = 4 \times 10^4 \text{ m s}^{-1}$ is a factor of 25 less than that for pristine graphene, $v_0 = 10^6 \text{ m s}^{-1}$. The measured oscillation frequencies indicate the existence of small Fermi pockets that start from the half-filling states, with half the degeneracy of the main Fermi surface of the Dirac points. Shaded regions at half-filling and full-filling correspond to the

shaded rectangles in Fig. 2a. The error bars in m^* and f_{SDH} give the uncertainty of fitting to the Lifshitz–Kosevich formula (defined in Methods) and correspond to the 90% confidence level. The blue dashed curves denote the f_{SDH} that is expected for Fermi surfaces with degeneracy $M = 4$ and $M = 2$, starting at charge neutrality and at the half-filling states, respectively. **c**, Gate dependence of the conductance of device D1 at different temperatures, 4.5 K, 8 K, 15 K, 29 K, 40 K, 50 K, 70 K, 100 K and 120 K. The curves are each shifted vertically by 0.006 mS for clarity. See Extended Data Fig. 5a, b for the temperature dependence up to room temperature. **d**, Comparison between the bandwidth W for the $E > 0$ flat-band branch in TBG (thick blue line) and the on-site energy $U = e^2 \theta / (4\pi \epsilon_0 \kappa a)$ (thin coloured lines for different values of κ) for different twist angles θ . Near the magic angles $\theta_{\text{magic}}^{(i)} \approx 1.1^\circ, 0.5^\circ, \dots$ for $i = 1, 2, \dots$, $U > W$ is satisfied for a range of possible values of κ (defined in the main text) and so the system can be driven into a Mott-like insulator state.

orientation suggests that the suppression of the half-filling states is due to a Zeeman effect rather than an orbital effect, because the latter would be affected by only the perpendicular component of the magnetic field. For an effective g -factor of $g = 2$ due to electron

spin, the Zeeman energy that is needed to suppress the half-filling states is approximately $g\mu_B B = 0.5 \text{ meV}$, where μ_B is the Bohr magneton—the same order of magnitude as that of the thermal excitation energy.

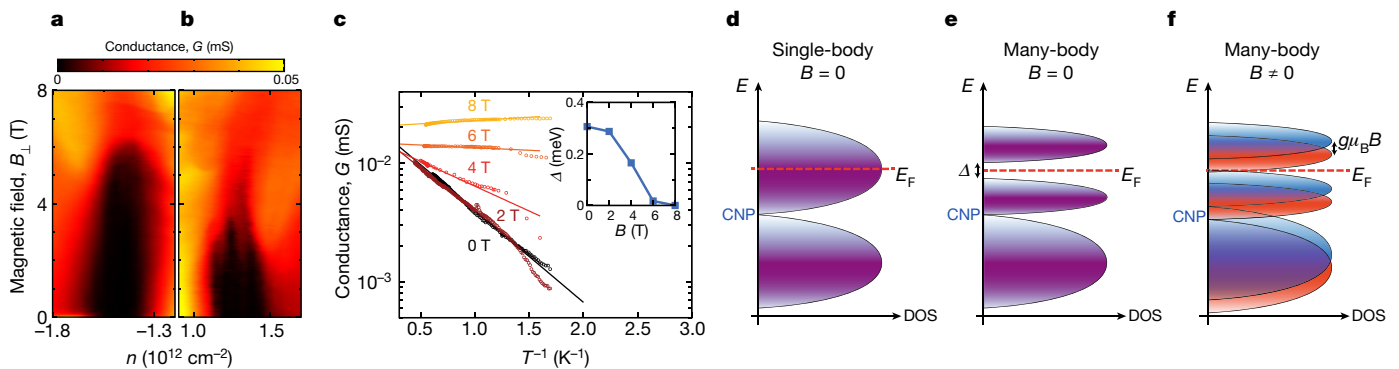


Figure 4 | Magnetic-field response of the half-filling insulating phases. **a**, **b**, Dependence of the conductance on the perpendicular magnetic field B_\perp of the half-filling states for device D1 on the p-side (**a**) and the n-side (**b**). The measurement is taken at 0.3 K. **c**, Arrhenius plot (circles) of the conductance of the p-side half-filling state at different magnetic fields. The inset shows the thermal activation gap Δ extracted from fitting the data in the main plot with $\exp[-\Delta/(2kT)]$ (solid lines). **d–f**, Schematics of the density of states (DOS) in different scenarios. The single-particle flat bands ($E > 0$ and $E < 0$ bands are both shown, with E_F in the $E > 0$

band (n -doping); **d**) are split into upper and lower many-body bands by interactions (**e**). This occurs when E_F is at half-filling of the upper band. Upon applying a Zeeman field ($B \neq 0$), the excitations can be further polarized, and can close the charge gap when the Zeeman energy $g\mu_B B$ is comparable to the gap Δ (**f**). Purple shading denotes a spin-degenerate band, whereas blue and red shading denotes spin-up and spin-down bands, respectively. CNP, charge neutrality point. The shape of the DOS drawn here is purely illustrative and does not represent the actual DOS profile (see Extended Data Fig. 6 for a numerical result).

Our data point to the presence of a spin-singlet Mott-like insulator ground state at half-filling and zero magnetic field (Fig. 4e). The application of an external magnetic field could polarize the excitations in the spectrum of the correlated states according to their spin. When the Zeeman energy exceeds the charge gap, charge conduction can therefore occur (Fig. 4f). In a typical Mott insulator, the ground state usually exhibits antiferromagnetic spin ordering below the Néel temperature. However, on a triangular lattice, the frustration prevents fully antiparallel alignment of adjacent spins. Possible ordering schemes include the 120° Néel order and a rotational-symmetry-breaking stripe order³⁰. It is unclear whether the spin-singlet ground state in magic-angle TBG is fulfilled by either of these ordering schemes, or whether it is disordered at low temperatures. In the half-filling states of magic-angle TBG it is also possible that the ordering, if any, occurs in conjunction with the valley degree of freedom. Therefore, any complete theoretical treatment of this problem should include a two-band Hubbard model on a triangular lattice.

We also comment on other competing mechanisms for creating a half-filled insulating state in a system with flat bands. Among the possibilities, charge-density waves in two dimensions are often stabilized by Fermi-surface nesting, which can in principle occur near the half-filling of a two-dimensional Brillouin zone³¹. However, this nesting is not sufficient to create a global gap in the entire Fermi surface to achieve an insulating state. To create a global gap at half-filling, at least a doubling of the unit cell would be necessary, which could be created by a commensurate charge-density wave or lattice relaxation due to strain. Scanning tunnelling microscopy conducted at temperatures below 4 K may be able to differentiate such mechanisms.

In summary, our work demonstrates that graphene can be transformed through van der Waals engineering into a flat-band system in which insulating states at half-filling are present. These insulating states cannot be explained in the absence of electron–electron interactions and so highlight the importance of correlations in this flat-band system. However, the lattice and electronic structure near magic-angle TBG superlattices is very complex, and further theoretical and experimental work is necessary to ascertain the importance of correlation effects fully. Through its easy gate tunability, magic-angle TBG could provide a way of studying the transition between a correlated metal and an interaction-driven insulating state, which could provide insights into strongly correlated materials, including high-temperature superconductivity. The combination of spin and valley degrees of freedom on a triangular lattice could also give rise to other exotic quantum phases, such as quantum spin liquids³².

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 October 2017; accepted 21 February 2018.

Published online 5 March 2018.

- Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).
- Hunt, B. *et al.* Massive Dirac fermions and Hofstadter butterfly in a van der Waals heterostructure. *Science* **340**, 1427–1430 (2013).
- Dean, C. R. *et al.* Hofstadter's butterfly and the fractal quantum Hall effect in moiré superlattices. *Nature* **497**, 598–602 (2013).
- Ponomarenko, L. A. *et al.* Cloning of Dirac fermions in graphene superlattices. *Nature* **497**, 594–597 (2013).
- Song, J. C. W., Shytov, A. V. & Levitov, L. S. Electron interactions and gap opening in graphene superlattices. *Phys. Rev. Lett.* **111**, 266801 (2013).
- Bistrizter, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
- Wu, C., Bergman, D., Balents, L. & Das Sarma, S. Flat bands and Wigner crystallization in the honeycomb optical lattice. *Phys. Rev. Lett.* **99**, 070401 (2007).
- Iglovikov, V. I., Hèbert, F., Grèmaud, B., Batrouni, G. G. & Scalettar, R. T. Superconducting transitions in flat-band systems. *Phys. Rev. B* **90**, 094506 (2014).

- Tsai, W. F., Fang, C., Yao, H. & Hu, J. Interaction-driven topological and nematic phases on the Lieb lattice. *New J. Phys.* **17**, 055016 (2015).
- Lieb, E. H. Two theorems on the Hubbard model. *Phys. Rev. Lett.* **62**, 1201–1204 (1989).
- Mielke, A. Exact ground states for the Hubbard model on the kagome lattice. *J. Phys. A* **25**, 4335–4345 (1992).
- Si, Q. & Steglich, F. Heavy fermions and quantum phase transitions. *Science* **329**, 1161–1166 (2010).
- Cao, Y. *et al.* Superlattice-induced insulating states and valley-protected orbits in twisted bilayer graphene. *Phys. Rev. Lett.* **117**, 116804 (2016).
- Suárez Morell, E., Correa, J. D., Vargas, P., Pacheco, M. & Barticevic, Z. Flat bands in slightly twisted bilayer graphene: tight-binding calculations. *Phys. Rev. B* **82**, 121407 (2010).
- Lopes dos Santos, J. M. B., Peres, N. M. R. & Castro Neto, A. H. Continuum model of the twisted graphene bilayer. *Phys. Rev. B* **86**, 155449 (2012).
- Fang, S. & Kaxiras, E. Electronic structure theory of weakly interacting bilayers. *Phys. Rev. B* **93**, 235153 (2016).
- Kim, K. *et al.* Tunable moiré bands and strong correlations in small-twist-angle bilayer graphene. *Proc. Natl Acad. Sci. USA* **114**, 3364–3369 (2017).
- Tramby de Laissardière, G., Mayou, D. & Magaud, L. Numerical studies of confined states in rotated bilayers of graphene. *Phys. Rev. B* **86**, 125413 (2012).
- Li, G. *et al.* Observation of van Hove singularities in twisted graphene layers. *Nat. Phys.* **6**, 109–113 (2010).
- Luican, A. *et al.* Single-layer behavior and its breakdown in twisted graphene layers. *Phys. Rev. Lett.* **106**, 126802 (2011).
- Brihuega, I. *et al.* Unraveling the intrinsic and robust nature of van Hove singularities in twisted bilayer graphene by scanning tunneling microscopy and theoretical analysis. *Phys. Rev. Lett.* **109**, 196802 (2012).
- Kim, K. *et al.* van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
- Ashoori, R. C. *et al.* Single-electron capacitance spectroscopy of discrete quantum levels. *Phys. Rev. Lett.* **68**, 3088–3091 (1992).
- Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004).
- Morozov, S. V. *et al.* Giant intrinsic carrier mobilities in graphene and its bilayer. *Phys. Rev. Lett.* **100**, 016602 (2008).
- Bolotin, K. I., Sikes, K. J., Hone, J., Stormer, H. L. & Kim, P. Temperature-dependent transport in suspended graphene. *Phys. Rev. Lett.* **101**, 096802 (2008).
- Mott, N. F. *Metal-Insulator Transitions* (Taylor and Francis, 1990).
- Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70**, 1039–1263 (1998).
- Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78**, 17–85 (2006).
- Misumi, K., Kaneko, T. & Ohta, Y. Mott transition and magnetism of the triangular-lattice Hubbard model with next-nearest-neighbor hopping. *Phys. Rev. B* **95**, 075124 (2017).
- Grüner, G. *Density Waves In Solids* (Westview Press, 2009).
- Balents, L. Spin liquids in frustrated magnets. *Nature* **464**, 199–208 (2010).

Acknowledgements We acknowledge discussions with L. Levitov, P. Lee, S. Todadri, B. I. Halperin, S. Carr, Z. Alpichshev, J. Y. Khoo and N. Staley. This work was primarily supported by the National Science Foundation (NSF; DMR-1405221) and the Gordon and Betty Moore Foundation's EPIQS Initiative through grant GBMF4541 for device fabrication, transport measurements and data analysis (Y.C., J.Y.L., J.D.S.-Y. and P.J.H.), with additional support from the NSS Program, Singapore (J.Y.L.). Capacitance work by R.C.A., A.D. and S.L.T. and theory work by S.F. was supported by the STC Center for Integrated Quantum Materials, NSF grant number DMR-1231319. Data analysis by V.F. was supported by AFOSR grant number FA9550-16-1-0382. K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by MEXT, Japan and JSPS KAKENHI grant numbers JP15K21722 and JP25106006. This work made use of the Materials Research Science and Engineering Center Shared Experimental Facilities supported by the NSF (DMR-0819762) and of Harvard's Center for Nanoscale Systems, supported by the NSF (ECS-0335765). E.K. acknowledges support by ARO MURI award W911NF-14-0247. R.C.A. acknowledges support by the Gordon and Betty Moore Foundation under grant number GBMF2931.

Author Contributions Y.C., J.Y.L. and J.D.S.-Y. fabricated the devices and performed transport measurements. Y.C. and V.F. performed data analysis. P.J.-H. supervised the project. S.F. and E.K. provided numerical calculations. S.L.T., A.D. and R.C.A. measured capacitance data. K.W. and T.T. provided hexagonal boron nitride devices. Y.C., V.F. and P.J.-H. wrote the paper with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.J.-H. (ppjarillo@mit.edu).

METHODS

Sample preparation. Devices D1, D2 and D4 were fabricated using a modified 'tear and stack' technique^{13,17,22}. Monolayer graphene and hexagonal boron nitride (10–30-nm thick) were exfoliated on SiO₂/Si chips and examined with optical microscopy and atomic force microscopy. We used a poly(bisphenol A carbonate) (PC)/polydimethylsiloxane (PDMS) stack on a glass slide mounted on a micro-positioning stage to first pick up a hexagonal boron nitride flake at 90 °C. Then we used the van der Waals force between hexagonal boron nitride and graphene to tear a graphene flake at room temperature. The separated graphene pieces were rotated manually by an angle θ about 0.2°–0.3° larger than the desired twist angle and stacked together again, resulting in a precisely controlled TBG structure. The TBG was then encapsulated by picking up another hexagonal boron nitride flake on the bottom, and the entire stack was released onto a metal gate at 160 °C. The final device geometry was defined by electron-beam lithography and reactive ion etching. Device D3 was fabricated using a slightly different procedure, whereby independent graphene flakes were stacked together. The edges of the graphene flakes were aligned under an optical microscope to obtain small twist angles.

Measurements. Transport measurements were performed using a standard low-frequency lock-in amplifier with an excitation frequency of about 10–20 Hz and an excitation voltage of 100 μ V, in a He-3 cryostat. The current flowing through the device was amplified by a current pre-amplifier and then measured by the lock-in amplifier.

The capacitance was measured using a low-temperature balanced capacitance bridge²³. A schematic of the measurement circuit is shown in Extended Data Fig. 2a. The reference capacitance C_{ref} used in our experiment was approximately 40 fF, and the device geometrical capacitance was approximately 7 fF. The a.c. excitation voltage used in our measurements was 3 mV at $f = 150$ kHz.

Transport data in device D4. Transport measurements in devices D1 and D3 were performed in a two-probe configuration. Although it is generally advised to perform four-probe measurements in transport experiments, we found that the existence of multiple insulating states (the superlattice gaps at $\pm n_s$ and the half-filling states at $\pm n_s/2$) frequently led to noisy or negative longitudinal resistance R_{xx} signals, owing to the region in the device near the voltage probes becoming insulating at a slightly different carrier density. In our case, where we are mostly interested in insulating behaviours of the order of 100 k Ω to 1 M Ω , a contact resistance of at most a few kilo-ohms, which is typical in edge-contacted graphene devices, did not obscure the data³³. We therefore believe that the two-probe data presented here can be trusted and provide an accurate representation of the device characteristics.

In Extended Data Fig. 4b, c we show the measurements of the two-probe and four-probe conductances in device D4, which has a twist angle of $\theta = 1.16^\circ \pm 0.02^\circ$. Device D4 was measured in a Hall-bar configuration so that the contact resistance could be removed. In this particular device, neither the superlattice insulating states nor the half-filling states had very high impedance (probably owing to disorder or inhomogeneity), and so the previously described issues with four-probe measurements did not occur. The four-probe and two-probe measurements show essentially the same features, although some weak signals appear to be better resolved in the four-probe measurements.

In the four-probe data, we not only observe the half-filling states (± 2 electrons per moiré unit cell), but we also see evidence for odd-filling insulating phases at ± 3 electrons per moiré unit cell, as a weak reduction in the conductance curve. The existence of insulating behaviours at integer fillings of the flat bands other than ± 2 is expected in Mott-like insulators and lends further support to our claim that the correlated insulating behaviour originates from the on-site Coulomb interaction.

Hall measurement in device D4. We also measured device D4 in a Hall configuration (transverse resistance R_{xy}). In Extended Data Fig. 4d, e we show the low-field linear Hall coefficient $R_H = R_{xy}/B$ and the Hall density $n_H = -1/(eR_H)$ versus the gate-induced charge density n . In a uniformly gated single-carrier two-dimensional electronic gas, we expect that $n_H = n$. This is indeed what we measured in the density range $-1.3 \times 10^{12} \text{ cm}^{-2}$ to $1.3 \times 10^{12} \text{ cm}^{-2}$ at 0.3 K. However, near the half-filling states $n = \pm n_s/2$, the Hall density jumps abruptly from $n_H = n$ to a small value close to zero (but without changing its sign). Beyond half-filling, n_H follows $n_H = n \pm n_s/2$, a trend that is consistent with quasiparticles that are generated from the half-filling states. This 'resetting' effect of the Hall density disappears gradually as the temperature is raised from 0.3 K to 10 K, in agreement with the energy scale of the Mott-like states. At higher temperatures, the Hall density is linear with n but the slope is no longer one, which might be related to the thermal energy kT being close to the bandwidth, which could result in thermally excited carriers with opposite polarity reducing the net Hall effect.

In good correspondence with the quantum oscillation data shown in Fig. 3b, we see the behaviours of the new quasiparticles on only one side of the Mott-like state, for example, the side farther from the charge neutrality point; between the

charge neutrality point and the Mott-like state, we see an abrupt change from the typical large Fermi surface of the single-particle bands to a small Fermi surface of the new quasiparticles. This may result if the effective mass of the quasiparticles on one side of the Mott-like gap is considerably greater than the other side, so that the oscillation and Hall effect become difficult to observe very close to the metal–insulator transition.

Determining the twist angle. Accurate determination of the twist angles of the samples is of utmost importance in understanding the magic-angle physics. We used several independent methods to determine the twist angle from the transport data.

First, the superlattice density n_s , defined as the density that is required to fill one band in the superlattice, is related to the twist angle by

$$n_s = \frac{4}{A} \approx \frac{8\theta^2}{\sqrt{3}a^2} \quad (1)$$

where A is the unit-cell area and $a = 0.246$ nm is the lattice constant of graphene. At approximately $1^\circ < \theta < 3^\circ$, the superlattice densities $\pm n_s$ are associated with a pair of single-particle bandgaps at their corresponding Fermi energy^{13,34,35}. The measured density of the insulating states of the superlattice can therefore be used to estimate θ directly according to equation (1). Owing to localized states, an accurate value of n_s is difficult to pinpoint at zero magnetic field, and the estimated θ has an uncertainty of about 0.1°–0.2°. In Extended Data Fig. 7a–d we show the resistivity (resistance for magic-angle device D1) for four different TBG samples with twist angles of $\theta = 1.38^\circ, 1.08^\circ, 0.75^\circ$ and 0.65° . At $\theta = 1.38^\circ$ and 1.08° , the positions of the superlattice gaps provide an good estimate of θ . However, it has been noted¹⁷ that the apparent resistance peaks in the transport data may not correspond to n_s but instead to $2n_s$, when the twist angle is below about 0.9° – 1° . We observed a similar phenomenon when the twist angle was as small as 0.65° . This complicates the determination of twist angles, because of the ambiguity of whether the feature observed corresponds to n_s or $2n_s$, which can result in the twist angle being wrong by a factor of $\sqrt{2}$.

Second, we use the fact that each band edge of the mini-band structure has its own Landau levels^{13,34,36}. In Extended Data Fig. 7e we show the magneto-conductance data of device D1 (first derivative with respect to n). The Landau levels emanating from $n_s = (2.7 \pm 0.1) \times 10^{12} \text{ cm}^{-2}$ can be clearly seen, which translates to $\theta = 1.08^\circ \pm 0.02^\circ$ according to equation (1). Because the intersection points of the Landau levels can be determined relatively accurately (uncertainty of about $1 \times 10^{11} \text{ cm}^{-2}$), the twist angle can be determined with an uncertainty of about 0.02° near the first magic angle.

Third, the effect of applying strong magnetic fields such that the magnetic length becomes comparable with the unit-cell size is described by Hofstadter's butterfly model³⁷. In density space, this model is better captured by Wannier³⁸. In the Wannier diagram, the Landau levels are universally represented by $n/n_s = \nu\phi/\phi_0 + s$, where ϕ is the magnetic flux through a unit cell, ν is an integer, and $s = 0$ labels the main Landau fan, $s = \pm 1$ is the first satellite fan, and so on. Adjacent Landau fans intersect when $\phi/\phi_0 = 1/q$ or, equivalently, $1/B = qA/\phi_0$, where q is another integer. Therefore, in the experiments we expect to see Landau-level crossings at periodic intervals of $1/B$, with the periodicity proportional to the unit-cell area A . This effect has been observed in other two-dimensional superlattice systems and can be used to cross-check the twist angles extracted from other methods^{2–4}. In Extended Data Fig. 7f we show the magneto-transport data (first derivative with respect to n) of device D3 at high doping densities, plotted versus n and $1/B$. A periodic crossing of Landau levels is observed near $-9 \times 10^{12} \text{ cm}^{-2}$ with period $0.033 \pm 0.001 \text{ T}^{-1}$, which gives $A = (1.37 \pm 0.04) \times 10^{-12} \text{ cm}^2$ and $\theta = 1.12^\circ \pm 0.01^\circ$, compared to $\theta = 1.12^\circ \pm 0.02^\circ$ extracted using the previous method ($n_s = (2.9 \pm 0.1) \times 10^{12} \text{ cm}^{-2}$).

Estimating the Fermi velocity from capacitance data. The measured capacitance is the series sum of the geometric capacitance C_{geom} and the quantum capacitance C_q . The latter is directly proportional to the DOS in TBG. Therefore, by analysing the quantum capacitance C_q as a function of carrier density n , we can extract the dependence of DOS on n and subsequently deduce the Fermi velocity.

In the zero-temperature limit, the quantum capacitance is related to the DOS $D(E)$ by $C_q = e^2 D(E_F)$, where E_F is the Fermi energy. In a model system for TBG near charge neutrality that consists of massless Dirac fermions with Fermi velocity v_F and eight-fold degeneracy (spin, valley, layer), the DOS is^{39–41}

$$D(E_F) = \frac{4}{\pi} E_F (\hbar v_F)^2$$

Because $E_F = \hbar v_F k_F$ (where k_F is the Fermi wavevector) is related to the density n by

$$n = 8 \frac{1}{(2\pi)^2} \pi k_F^2 = \frac{2}{\pi} \frac{E_F^2}{(\hbar v_F)^2}, \quad E_F = \hbar v_F \sqrt{\frac{n\pi}{2}}$$

where the factor of 8 comes from the spin, valley and layer degeneracy, the quantum capacitance of the TBG is

$$C_q = e^2 \frac{2\sqrt{2}}{\sqrt{\pi} \hbar v_F} \sqrt{|n| + n_d} \quad (2)$$

Owing to disorder, the spatially averaged DOS at the Dirac point ($n = E_F = 0$) will not be absolutely zero. Therefore, a phenomenological disorder density of $n_d \approx 1 \times 10^{10} \text{ cm}^{-2}$ is added in the above expression³⁹.

The measured capacitance is then

$$\frac{1}{C} = \frac{1}{C_{\text{geom}}} + \frac{1}{C_q} \quad (3)$$

In Extended Data Fig. 2b, we show the measured capacitance near the Dirac point and fitting curves according to equations (2) and (3). C_{geom} is approximated by the d.c. gating capacitance $C_g \approx 7.5 \text{ fF}$. We find that using $v_F = 0.15 \times 10^6 \text{ m s}^{-1}$ and $n_d = 1 \times 10^{10} \text{ cm}^{-2}$ gives a reasonable fit to the data measured at both 0.3 K and 2 K.

The fitting for v_F is sensitive to the value used for C_{geom} . For example, using a C_{geom} value 30% larger than the value that we used above, we find a Fermi velocity of $v_F = 0.10 \times 10^6 \text{ m s}^{-1}$. Similarly, using a value 15% smaller than the said value we find $v_F = 0.20 \times 10^6 \text{ m s}^{-1}$. Nonetheless, the analysis presented here suffices to demonstrate that the Fermi velocity is indeed reduced greatly in the capacitance device D2. The slightly larger Fermi velocity compared to that measured in the transport device D1 ($v_F = 0.04 \times 10^6 \text{ m s}^{-1}$) can be attributed to the slightly larger twist angle of device D2 ($\theta = 1.10^\circ$), which might be farther from the first magic angle $\theta_{\text{magic}}^{(1)} \approx 1.05^\circ$.

Error bars. The error bars in the inset of Fig. 2a are computed using the following criteria: for the transport devices D1, D3 and D4, the endpoints of the error bars correspond to the points at which the conductance rises to 10% of the peak value on that side; for the capacitance device D2, because the peaks are very sharp (see Fig. 3a), the error bar corresponds to the width of the entire peak in the loss tangent data.

Quantum oscillations and extracting m^* . We performed magneto-transport measurements in device D1 from 0.3 K to 10 K. At each gate voltage, a polynomial background of resistance in B was first removed, and then the oscillation frequency and the effective mass was analysed. Examples of the Shubnikov–de Haas oscillations and their temperature dependences at a few representative gate voltages are shown in Extended Data Fig. 3a–c. The temperature dependence of the most prominent peak is fitted with the Lifshitz–Kosevich formula applied to the resistance:

$$\Delta R \propto \frac{\chi}{\sinh(\chi)}, \quad \chi = \frac{2\pi^2 k T m^*}{\hbar e B} \quad (4)$$

where ΔR is the change in resistance and the cyclotron mass m^* is extracted from the fitting (examples shown in Extended Data Fig. 3d). Within the flat bands, the quantum oscillations universally disappear at around 10 K except very close to the Dirac point, consistent with the large electron mass and greatly reduced Fermi velocity near the first magic angle.

The full magneto-conductance map measured in device D1 at 0.3 K is shown in Extended Data Fig. 5c. At first glance, it may seem that the Landau levels that emanate from the Dirac point ‘penetrate’ the half-filling states and continue towards the band edges. However, this is not the case. In Extended Data Fig. 3e, f we show the same data but plotted versus $1/B$ instead of B . Here it can be seen that at densities beyond the half-filling states, the oscillations clearly do not converge at the Dirac point, instead converging at the half-filling states. The oscillation frequencies extracted from these data are plotted in Fig. 3b.

Band structure of TBG near magic angles. The general evolution of the band structure of TBG above the first magic angle has been described previously^{6,14–16,18,34,35}. The low-energy band structure consists of two Dirac cones (each is four-fold-degenerate owing to valley and spin), with a renormalized Fermi velocity of

$$v_F(\theta) = v_0 \frac{1 - 3\alpha^2}{1 + 6\alpha^2}$$

where $\alpha = w/(\hbar v_0 k_\theta)$ is the dimensionless interlayer hopping amplitude (w and v_0 are the interlayer hopping energy and original Fermi velocity in graphene, $k_\theta \approx G_K \theta$ is the interlayer momentum difference and G_K is the wave number at the corner of the Brillouin zone in graphene)^{6,15}. When $\alpha \ll 1$, $v_F(\theta)$ is approximately $v_0(1 - 9\alpha^2)$. $v_F(\theta)$ passes through zero at $\alpha = 1/\sqrt{3}$, which defines the first magic angle $\theta_{\text{magic}}^{(1)}$. However, to our knowledge, the detailed evolution of the band structure near magic angles has not been addressed previously. Specifically, we seek to

determine how the associated winding number evolves as the Fermi velocity at the Dirac points changes sign. Close to a generic Dirac point, the effective two-band Hamiltonian can be written as⁴²

$$\mathcal{H}(\mathbf{k}) = \hbar v_F(\theta) \boldsymbol{\sigma} \cdot \mathbf{k} + \mathcal{O}(k^2) = \begin{bmatrix} \mathcal{O}(k^2) & \hbar v_F(\theta) k^\dagger + \mathcal{O}(k^2) \\ \hbar v_F(\theta) k + \mathcal{O}(k^2) & \mathcal{O}(k^2) \end{bmatrix}$$

in which $\mathbf{k} = (k_x, k_y)$, $k = k_x + ik_y$ and $\boldsymbol{\sigma} = (\sigma_x, \sigma_y)$ is the vector of the Pauli matrices. As $v_F(\theta) \rightarrow 0$ near the first magic angle, the terms that are linear in k vanish and the dispersion is dominated by the next-leading-order, k^2 terms. A simple form for the k^2 term is

$$\mathcal{H}(\mathbf{k}) = \begin{bmatrix} 0 & \hbar v_F(\theta) k^\dagger + \frac{\hbar^2}{2m} k^2 \\ \hbar v_F(\theta) k + \frac{\hbar^2}{2m} (k^\dagger)^2 & 0 \end{bmatrix} \quad (5)$$

in which m is a parameter with the dimension of mass. This Hamiltonian describes the low-energy band dispersion of monolayer graphene with third-nearest-neighbour hopping and of bilayer graphene with Bernal stacking and trigonal warping^{42–46}. The eigenvalues of this Hamiltonian are

$$E_{\pm}(\mathbf{k}) = \pm \sqrt{\left[\hbar v_F k_x + \frac{\hbar^2}{2m} (k_x^2 - k_y^2) \right]^2 + \left[\hbar v_F k_y - \frac{\hbar^2}{m} k_x k_y \right]^2} \quad (6)$$

The evolution of the dispersion described by equation (6) with varying v_F and a constant $m = 0.5$ is shown in Extended Data Fig. 1a–f. The winding number associated with a Dirac point is defined by

$$w = \frac{i}{2\pi} \oint_C \langle \mathbf{k} | \nabla_{\mathbf{k}} | \mathbf{k} \rangle d\mathbf{k}$$

where C is a loop around the Dirac point. The winding number follows a conservation law when the motion and merging of Dirac points are considered⁴². The winding number of the upper band ($E_+(\mathbf{k})$) at each touching point in the upper band is labelled in Extended Data Fig. 1a–f.

As $v_F \rightarrow 0$ there exist three additional Dirac points with opposite winding number (-1) to the main Dirac point ($+1$). Therefore, at $v_F = 0$, when all four Dirac points merge, the winding number is -2 because the total winding number cannot change.

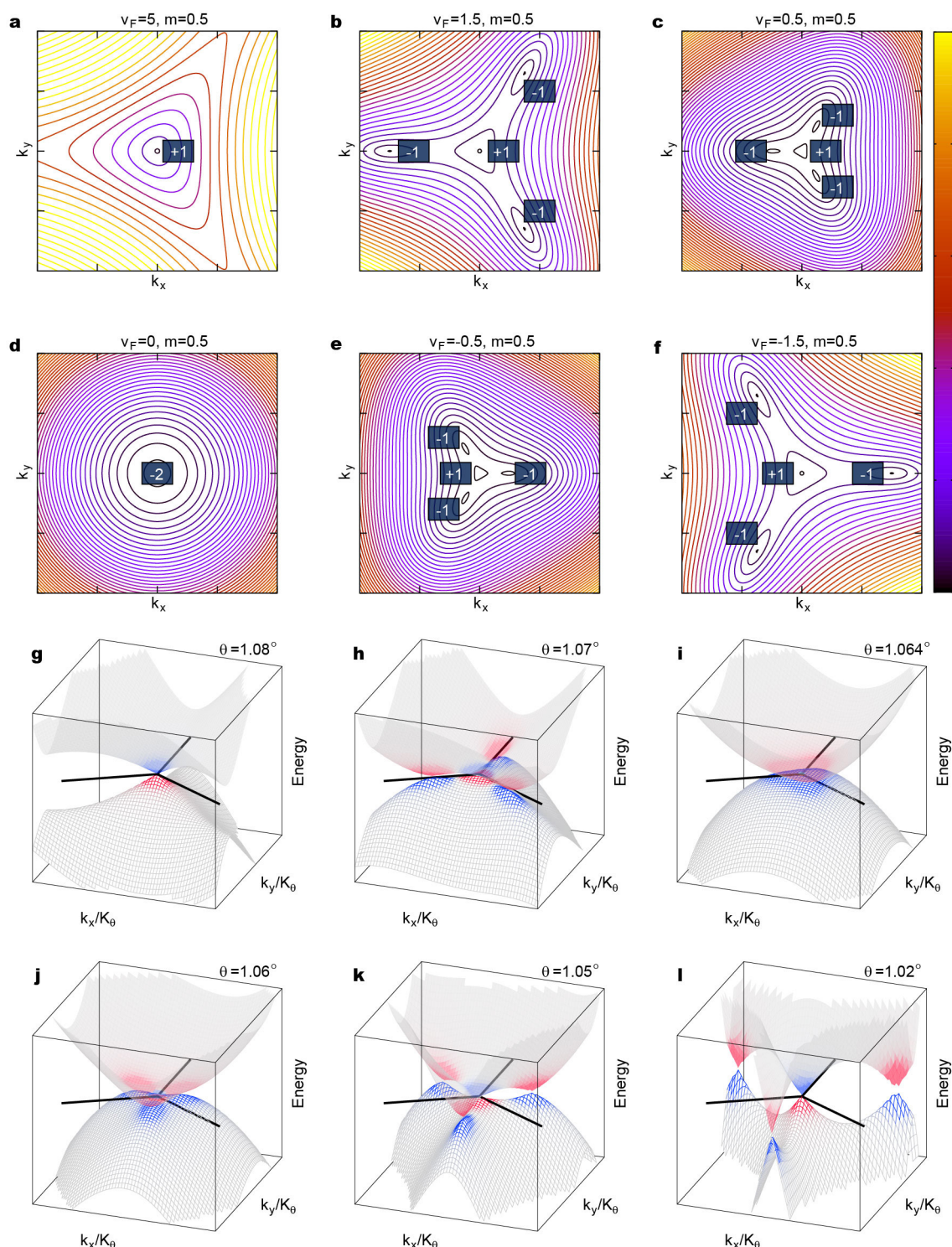
The simple Hamiltonian form in equation (5) is an educated guess. We performed numerical calculations of the winding number using the continuum model for TBG^{6,15} and the numerical method in ref. 47. The results are summarized in Extended Data Fig. 1g–l. We find that near the first magic angle of the model used, $\theta_{\text{magic}}^{(1)} = 1.064^\circ$, the behaviour described in Extended Data Fig. 1a–f is exactly what happens at each corner of the mini Brillouin zone. The complication that arises when we consider the entire mini Brillouin zone is that, for a given valley (of the original graphene Brillouin zone, such as K), the two inequivalent corners of the mini Brillouin zone have the same winding number because they are the hybridized result of the same valley (K) of opposite layers (see Fig. 1d). Global time-reversal symmetry is preserved by mapping to the other valley (K'). Therefore, for a given valley K , when the twist angle is reduced from large angles, at which the winding numbers of the two corners are $(+1, +1)$, to the first magic angle, at which the winding numbers are $(-2, -2)$, a net winding number change of $\Delta w = 6$ occurs between the two lowest-energy bands. Further theoretical work is necessary to elucidate the physics behind this winding number evolution near the first magic angle.

In summary, we show that at exactly the first magic angle the Dirac point at each corner of the mini Brillouin zone (K_s and K'_s) becomes a parabolic band touching with a winding number of -2 , similarly to bilayer graphene with Bernal stacking except that the two corners have the same winding number. The calculation that corresponds to the first magic angle in Extended Data Fig. 1i can be fitted to a paraboloid, which yields an effective mass of $1.1 m_e$. This value can be viewed as the asymptotic limit of the effective mass near the charge neutrality point as $v_F \rightarrow 0$. **DOS in magic-angle TBG.** Despite our simplistic representation of the DOS in the flat bands of magic-angle TBG (Fig. 4d–f), the actual single-particle DOS profile of magic-angle TBG is more complex, with multiple van Hove singularities. In Extended Data Fig. 6 we show the DOS versus energy calculated using a continuum model⁶ for $\theta = 1.08^\circ$.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

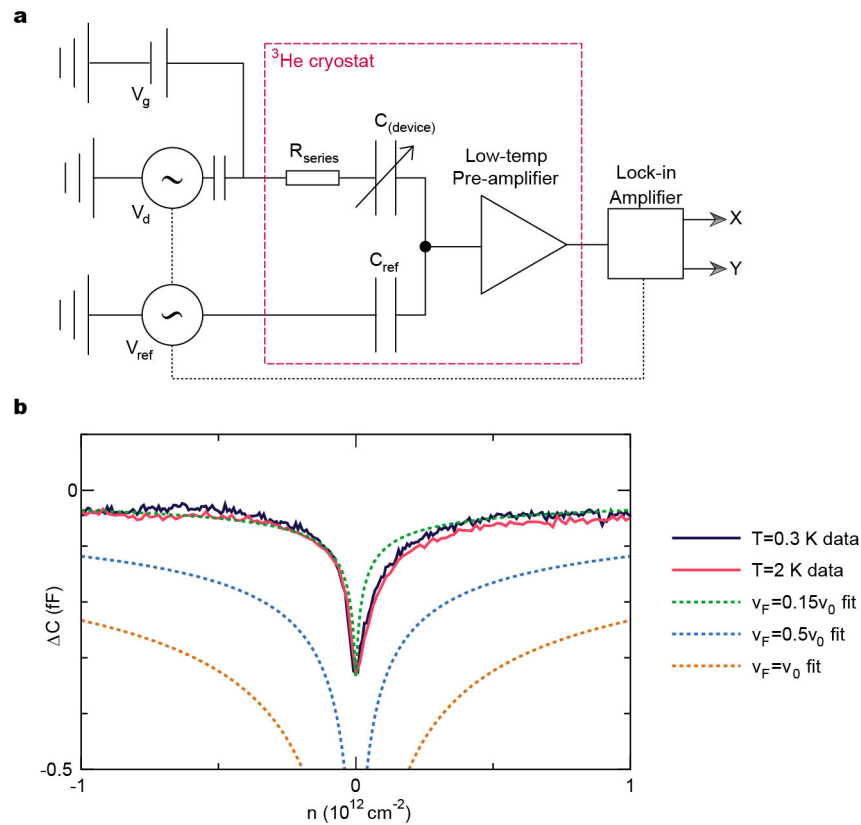
33. Wang, L. *et al.* One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).

34. Moon, P. & Koshino, M. Energy spectrum and quantum Hall effect in twisted bilayer graphene. *Phys. Rev. B* **85**, 195458 (2012).
35. Nam, N. N. T. & Koshino, M. Lattice relaxation and energy band modulation in twisted bilayer graphenes. *Phys. Rev. B* **96**, 075311 (2017).
36. Kim, Y. *et al.* Charge inversion and topological phase transition at a twist angle induced van Hove singularity of bilayer graphene. *Nano Lett.* **16**, 5053–5059 (2016).
37. Hofstadter, D. R. Energy levels and wave functions of Bloch electrons in rational and irrational magnetic fields. *Phys. Rev. B* **14**, 2239–2249 (1976).
38. Wannier, G. H., A result not dependent on rationality for Bloch electrons in a magnetic field. *Phys. Status Solidi b* **88**, 757–765 (1978).
39. Xia, J., Chen, F., Li, J. & Tao, N. Measurement of the quantum capacitance of graphene. *Nat. Nanotechnol.* **4**, 505–509 (2009).
40. Fang, T., Aniruddha, K., Xing, H. & Jena, D. Carrier statistics and quantum capacitance of graphene sheets and ribbons. *Appl. Phys. Lett.* **91**, 092109 (2007).
41. Wallace, P. R. The band theory of graphite. *Phys. Rev.* **71**, 622–634 (1947).
42. Goerbig, M. & Montambaux, G. in *Dirac Matter* (eds Duplantier, B. *et al.*) 25–53 (Springer, 2017).
43. Bena, C. & Simon, L. Dirac point metamorphosis from third-neighbor couplings in graphene and related materials. *Phys. Rev. B* **83**, 115404 (2011).
44. Montambaux, G. An equivalence between monolayer and bilayer honeycomb lattices. *Eur. Phys. J. B* **85**, 375 (2012).
45. McCann, E. & Koshino, M. The electronic properties of bilayer graphene. *Rep. Prog. Phys.* **76**, 056503 (2013).
46. McCann, E. & Fal'ko, V. I. Landau-level degeneracy and quantum Hall effect in a graphite bilayer. *Phys. Rev. Lett.* **96**, 086805 (2006).
47. Fukui, T., Hatsugi, Y. & Suzuki, H. Chern numbers in discretized Brillouin zone: efficient method of computing (spin) Hall conductances. *J. Phys. Soc. Jpn* **74**, 1674–1677 (2005).



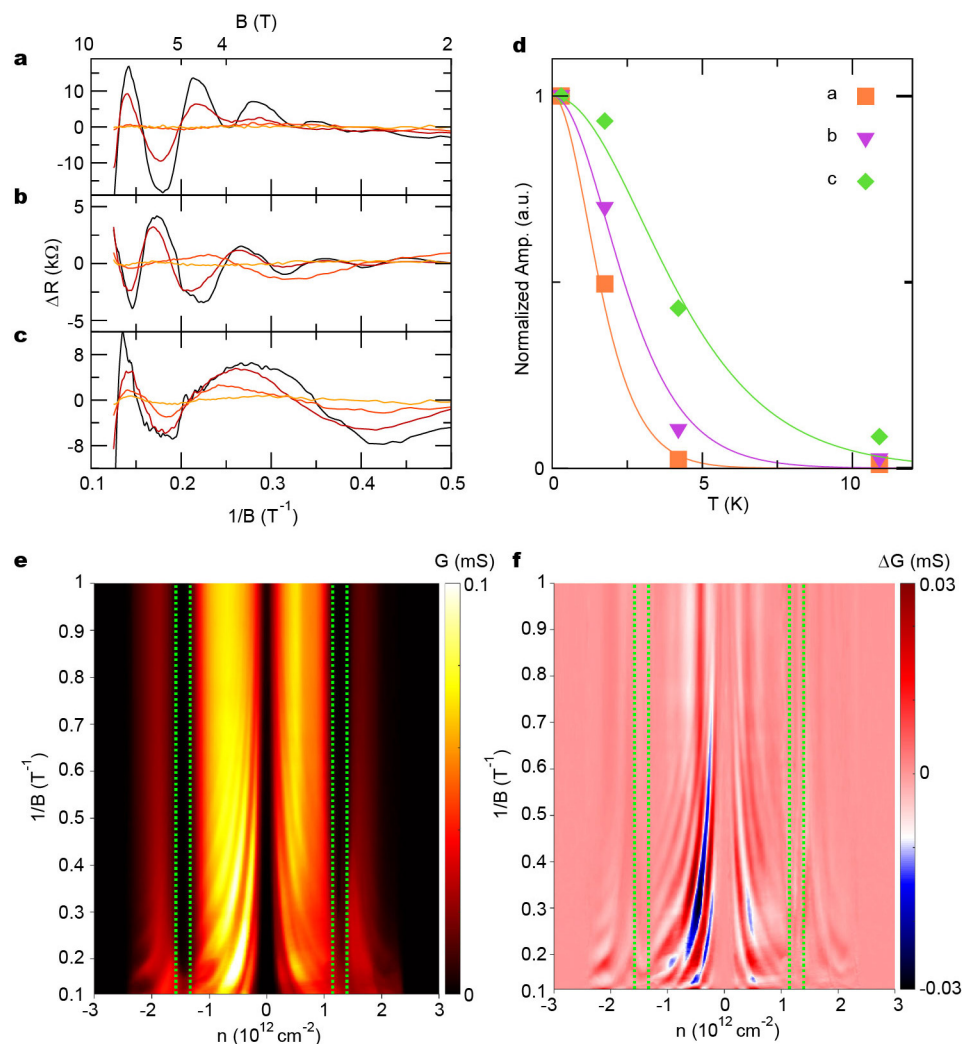
Extended Data Figure 1 | Evolution of the low-energy band structure of TBG near the magic angle. a–f, E_+ dispersion as in equation (6) for different v_F and fixed $m=0.5$. The k_x and k_y range in the figures is $[-2, 2]$ and the colour scale (on the right side of the figures) for the dimensionless energy axis is 0 to 10 from bottom to top. The associated winding number of each touching point is labelled. g–i, The evolution of the low-energy band structure of TBG near the first magic angle $\theta_{\text{magic}}^{(1)} = 1.064^\circ$ in the model. The colour shows the hotspots of the Berry curvature at the

touching point of each band. The energy axis spans an extremely small range of $[-50, 50] \mu\text{eV}$. The momentum axes are measured by $k_\theta \approx K\theta$ and the range for both k_x/K_θ and k_y/K_θ is $[-0.1, 0.1]$. The centre of the momentum space is the K_s point of the mini Brillouin zone (see Fig. 1d), and the thick lines denotes the K_s – M_s – K'_s directions (there are three inequivalent ones). All results are shown for the K-valley continuum description of TBG⁶.



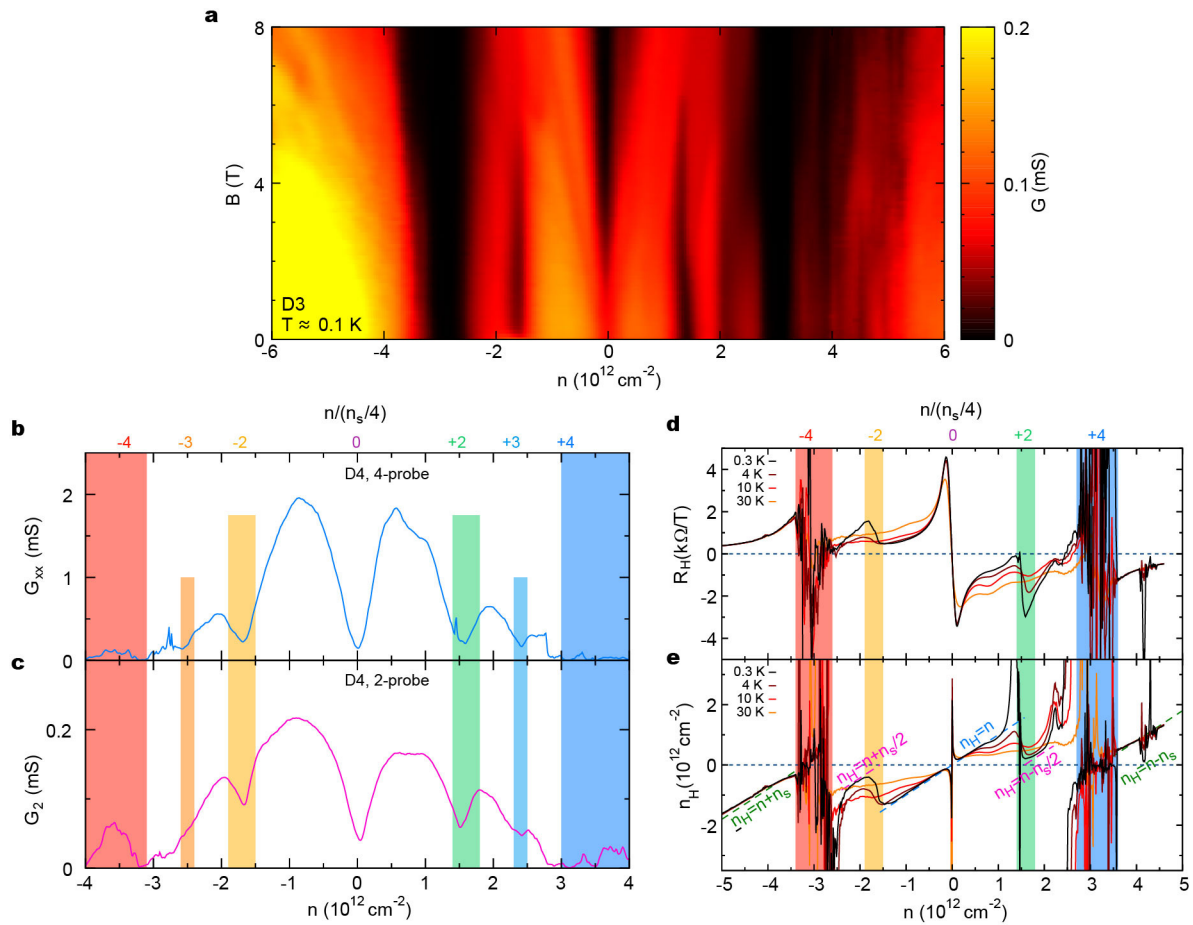
Extended Data Figure 2 | Capacitance measurement set-up and extraction of the Fermi velocity. **a**, Schematic of the low-temperature capacitance bridge. The X and Y outputs from the lock-in amplifier refer to the in-phase and out-of-phase components, respectively. $C_{\text{(device)}}$ and R_{series} are the capacitance and resistance of the sample. V_g is the d.c. gate voltage, V_d is the excitation voltage, V_{ref} is the reference voltage and C_{ref} is

the reference capacitance. All connections into and out of the cryostat are made with coaxial cables. **b**, Capacitance ΔC of device D2 near the charge neutrality point, and fitting curves according to equations (4) and (5) with different Fermi velocities. $v_0 = 10^6 \text{ m s}^{-1}$ is the Fermi velocity in pristine graphene.



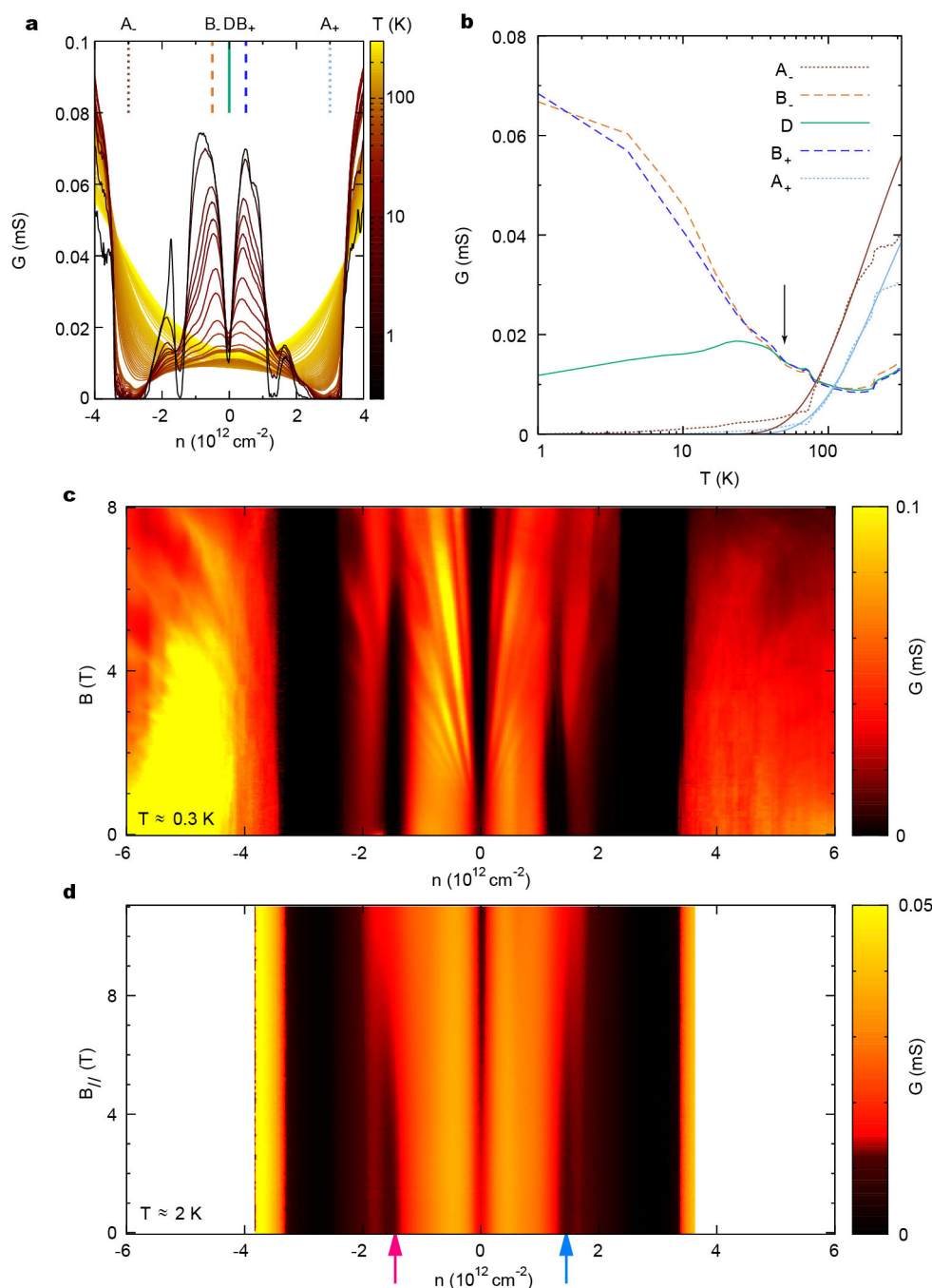
Extended Data Figure 3 | Quantum oscillations and extraction of the effective mass. **a–c**, Temperature-dependent magneto-resistance ΔR of device D1 at gate-voltage-induced carrier densities of $n = -2.08 \times 10^{12} \text{ cm}^{-2}$ (a), $n = -1.00 \times 10^{12} \text{ cm}^{-2}$ (b) and $n = 0.19 \times 10^{12} \text{ cm}^{-2}$ (c). The temperatures are, from dark to bright, 0.3 K, 1.7 K, 4.2 K and 10.7 K. **d**, Oscillation amplitudes of the most prominent peaks in **a–c**. The curves are fitted according to the Lifshitz–Kosevich formula

(equation (4)). **e**, Magneto-conductance G of device D1 (measured at 0.3 K) plotted versus n and $1/B$. **f**, The same data with a polynomial background in B removed for each density. The green boxes denote the range of densities for the half-filling states. At densities beyond the half-filling states, the oscillations do not converge at the Dirac point, but instead at the half-filling states.



Extended Data Figure 4 | Supplementary transport data in devices D3 and D4. **a**, Magneto-conductance G in device D3 ($\theta = 1.12^\circ$) versus n and B . The primary features at the superlattice gaps $\pm n_s$ and the half-filling states $\pm n_s/2$ are essentially identical to those for device D1. **b**, **c**, Four-probe (**b**; G_{xx}) and two-probe (**c**; G_2) conductance measured in device D4 ($\theta = 1.16^\circ$) at 0.3 K. The coloured vertical bars and the corresponding numbers indicate the associated integer filling inside each unit cell of the moiré pattern. As well as the half-filling states (± 2), we also observe weak

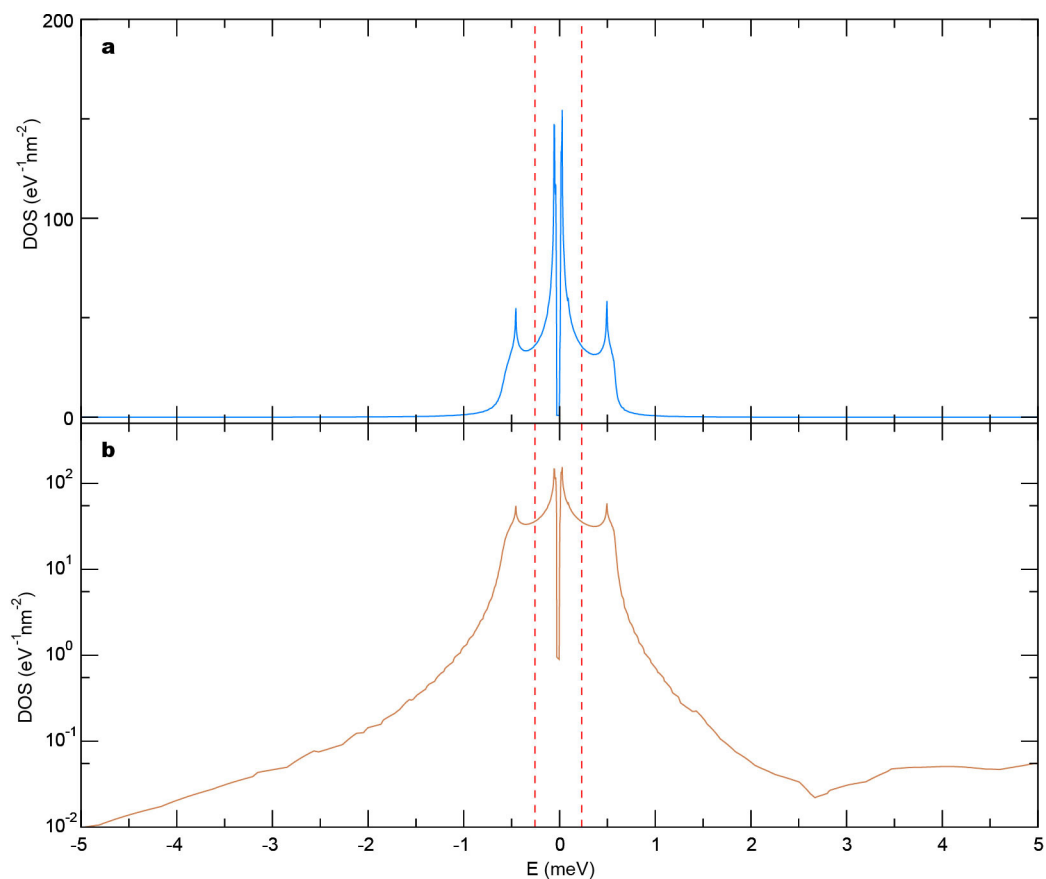
drops in the four-probe conductivity that point to three-quarter-filling states at ± 3 . **d**, **e**, Hall measurement in device D4 at various temperatures: the Hall coefficients R_H (**d**) and the Hall density $n_H = -1/(eR_H)$ (**e**). The coloured vertical bars and the corresponding numbers are as in **b** and **c**. The x axis is the gate-induced total charge density n , whereas the Hall density n_H and its sign indicate the number density and characteristic (electron-like or hole-like) curve of the carriers being transported.



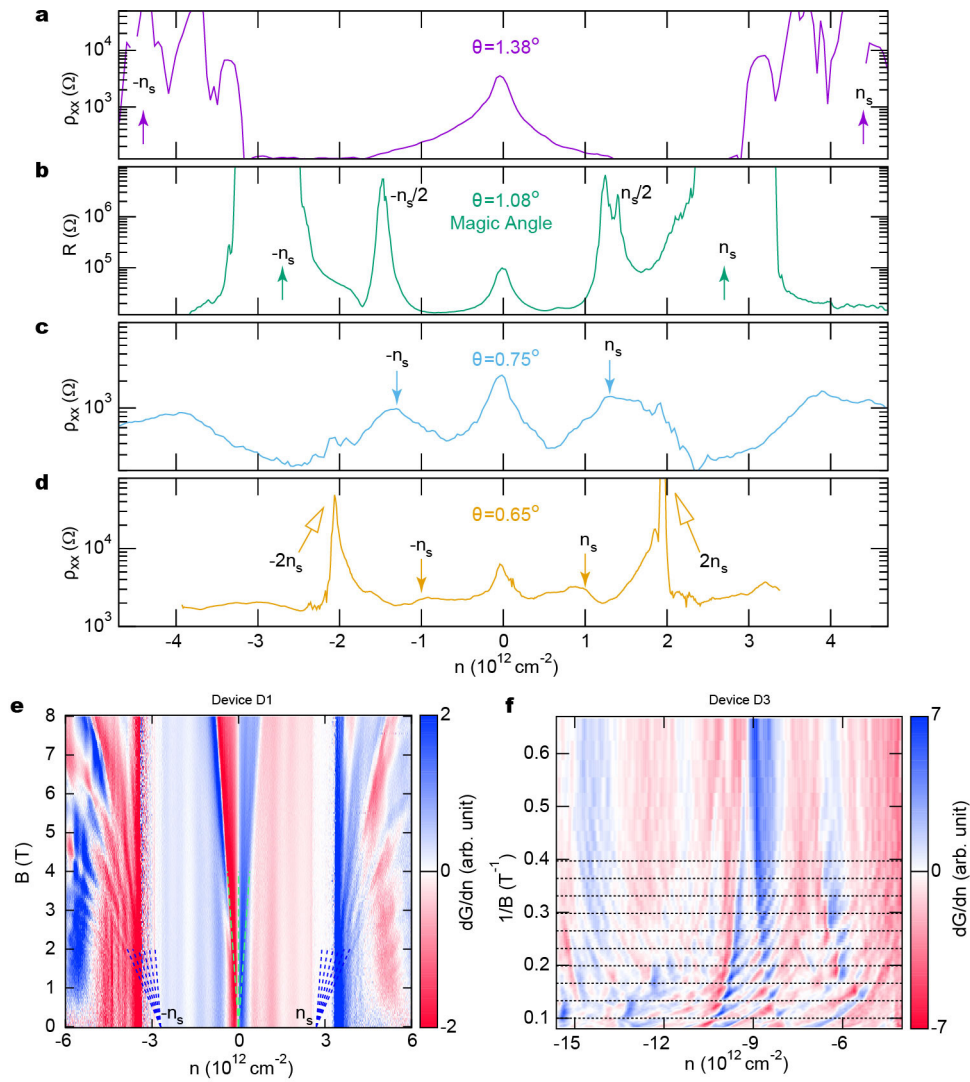
Extended Data Figure 5 | Supplementary transport data in device D1.

a, Temperature dependence of the conductance G of device D1 from 0.3 K to 300 K. **b**, The conductance versus temperature at five characteristic carrier densities, labelled A_{\pm} (superlattice gaps), B_{\pm} (above and below the Dirac point) and D (the Dirac point) in **a**. The arrow denotes the temperature above which the conductances at B_{\pm} merge with that of D . The solid lines accompanying the A_{\pm} traces are Arrhenius fits to the data. The thermal activation gaps of the superlattice insulating states at A_{\pm} can be obtained by fitting the temperature dependence of the conductance at these densities. See ref. 13 for a detailed discussion about the superlattice gaps in non-magic-angle devices. The fit to the Arrhenius formula $\exp[-\Delta/(2kT)]$ yields $\Delta_- = 32 \text{ meV}$ for the A_- gap and $\Delta_+ = 40 \text{ meV}$ for the A_+ gap. For comparison, the same gaps measured in $\theta = 1.8^\circ$ TBG are

slightly larger, $\Delta_- = 50 \text{ meV}$ and $\Delta_+ = 60 \text{ meV}$ for the gaps at negative and positive densities, respectively¹³. **c**, Magneto-conductance in device D1 as a function of gate-induced charge density n and perpendicular magnetic field B_{\perp} . **d**, Magneto-conductance in device D1 measured as a function of n and in-plane magnetic field B_{\parallel} . The in-plane measurement is made at a higher temperature of about 2 K. Combined with the degradation of the sample quality that resulted from the thermal cycling that was necessary to change the field orientation, the half-filling states are not as well developed as in the previous measurements. However, the gradual suppression of the half-filling states is still unambiguously observed when B_{\parallel} is above about 6 T, slightly higher but similar to the approximately 4–6-T threshold for the perpendicular field (see **c** and Fig. 4a, b). The red and blue arrows point to the p-side and n-side half-filling states, respectively.



Extended Data Figure 6 | DOS in magic-angle TBG. **a, b**, Single-particle DOS in TBG at $\theta = 1.08^\circ$, on linear (**a**) and logarithmic (**b**) scales. The red dashed lines denote the energy at which the lower and upper flat bands are half-filled. The results are obtained numerically using a continuum model⁶.



Extended Data Figure 7 | Determining the twist angle. **a–d**, Resistivity ρ_{xx} (resistance R for the $\theta = 1.08^\circ$ device) measurements for four samples with different twist angles: $\theta = 1.38^\circ$ (**a**), $\theta = 1.08^\circ$ (**b**), $\theta = 0.75^\circ$ (**c**) and $\theta = 0.65^\circ$ (**d**). The filled arrows highlight superlattice features at $\pm n_s$ and open arrows highlight $\pm 2n_s$ features that may correspond to features reported in ref. 17. So far, we have observed the half-filling states only in devices that have twist angles within 0.1° of the first magic angle. **e**, Magneto-conductance data (derivative with respect to n ; dG/dn) of

device D1 ($\theta = 1.08^\circ$) measured at 4 K. The dashed lines label the main (green) and satellite (blue) Landau fans. From the convergence point of the blue fans, we can accurately determine the superlattice density n_s and thus θ , with an uncertainty of about 0.02° . **f**, Hofstadter's oscillation manifested as periodic crossings of Landau levels in $1/B$. Data shown is the magneto-conductance (derivative with respect to n ; dG/dn) of device D3 ($\theta = 1.12^\circ$). The horizontal lines have a uniform spacing of $0.033 \pm 0.001 \text{ T}^{-1}$, which corresponds to $\theta = 1.12^\circ \pm 0.01^\circ$.

Room-temperature nine- μm -wavelength photo-detectors and GHz-frequency heterodyne receivers

Daniele Palaferri¹, Yanko Todorov¹, Azzurra Biglioli¹, Alireza Mottaghizadeh¹, Djamal Gacemi¹, Allegra Calabrese¹, Angela Vasanelli¹, Lianhe Li², A. Giles Davies², Edmund H. Linfield², Filippas Kapsalidis³, Mattias Beck³, Jérôme Faist³ & Carlo Sirtori¹

Room-temperature operation is essential for any optoelectronics technology that aims to provide low-cost, compact systems for widespread applications. A recent technological advance in this direction is bolometric detection for thermal imaging¹, which has achieved relatively high sensitivity and video rates (about 60 hertz) at room temperature. However, owing to thermally induced dark current, room-temperature operation is still a great challenge for semiconductor photodetectors targeting the wavelength band between 8 and 12 micrometres², and all relevant applications, such as imaging, environmental remote sensing and laser-based free-space communication^{3–5}, have been realized at low temperatures. For these devices, high sensitivity and high speed have never been compatible with high-temperature operation^{6,7}. Here we show that a long-wavelength (nine micrometres) infrared quantum-well photodetector⁸ fabricated from a metamaterial made of sub-wavelength metallic resonators^{9–12} exhibits strongly enhanced performance with respect to the state of the art up to room temperature. This occurs because the photonic collection area of each resonator is much larger than its electrical area, thus substantially reducing the dark current of the device¹³. Furthermore, we show that our photonic architecture overcomes intrinsic limitations of the material, such as the drop of the electronic drift velocity with temperature^{14,15}, which constrains conventional geometries at cryogenic operation⁶. Finally, the reduced physical area of the device and its increased responsivity allow us to take advantage of the intrinsic high-frequency response of the quantum detector⁷ at room temperature. By mixing the frequencies of two quantum-cascade lasers¹⁶ on the detector, which acts as a heterodyne receiver, we have measured a high-frequency signal, above four gigahertz (GHz). Therefore, these wide-band uncooled detectors could benefit technologies such as high-speed (gigabits per second) multichannel coherent data transfer¹⁷ and high-precision molecular spectroscopy¹⁸.

An unexploited intrinsic property of inter-subband quantum-well infrared photodetectors (QWIPs) based on group III–V semiconductor materials is the very short lifetime of their excited carriers. The typical lifetime is of the order of a few picoseconds⁷, which has two important consequences: the detector frequency response can reach 100 GHz and its saturation intensity is very high¹⁹ (10^7 W cm^{-2}). These properties are ideal for a heterodyne detection scheme in which a powerful local oscillator can drive a strong photocurrent (higher than the dark current of the detector) that can coherently mix with a signal shifted in frequency with respect to the local oscillator. Notably, these properties are unobtainable in infrared inter-band detectors based on mercury–cadmium–telluride (MCT) alloys, which have a much longer carrier lifetime and therefore a lower-speed response^{2,20,21}. However, the performance of all photonic detectors is limited by their high dark current, which originates from thermal emission of electrons from the wells and rises exponentially with temperature, thus requiring

cryogenic operation (about 80 K) for high-sensitivity measurements. Highly doped²² (about 10^{12} cm^{-2}), photovoltaic²³ 10- μm -wavelength QWIPs and quantum cascade detectors²⁴ with a large number of quantum wells have been observed to operate up to room temperature, but only when illuminated with powerful sources, such as CO₂ or free-electron lasers.

In this work, we show that this intrinsic limitation in QWIP detectors can be overcome through the use of a photonic metamaterial. We are able to calibrate our detector at room temperature using a blackbody emitting energy of only hundreds of nanowatts, orders of magnitude smaller than that required previously. Until now, room-temperature performance comparable with that reported here has only been demonstrated in the 3–5 μm wavelength range, using quantum cascade detectors^{24–26} and standard MCT detectors²⁷.

The photonic metamaterial structure of our detectors is shown in Fig. 1a. The GaAs/AlGaAs QWIP⁸ contains $N_{\text{qw}} = 5$ quantum wells absorbing at a wavelength of 8.9 μm (139 meV) and has been designed according to an optimized bound-to-continuum structure described in ref. 7. The absorbing region is inserted in an array of double-metal patch resonators^{9–12}, which provide sub-wavelength electric field confinement and act as antennas. The resonance wavelength is defined by the patch size s according to $\lambda = 2sn_{\text{eff}}$, where $n_{\text{eff}} = 3.3$ is the effective index⁹. The structures with $s = 1.3 \mu\text{m}$ are therefore in resonance with the peak responsivity of the detector.

In our structure, the microcavity increases the responsivity of the device by enhancing the local field in the thin semiconductor absorber¹⁰. In addition, the antenna effect extends the photon collection area of the detector, A_{coll} , making it much larger than the electrical area $\sigma = s^2$ of the device¹³. Because the detector photocurrent is proportional to A_{coll} , whereas the dark current is proportional to σ , for the same number of collected photons there is a substantial reduction of the dark current with respect to the case $\sigma = A_{\text{coll}}$, which results in a net increase of the operating temperature of the detector.

Besides the collection area A_{coll} , which defines the absorption cross-section per patch resonator, another crucial parameter is the contrast C of the reflectivity resonance (Fig. 1b). This parameter quantifies the fraction of the incident photon flux that is absorbed collectively by the array. As shown in Fig. 1c, the contrast can be adjusted¹⁰ by changing the array periodicity p . Optimal detector responsivity is obtained at the critical coupling point, $C = 1$, where all incident radiation is coupled into the array. The collection area per patch is related to the contrast according to the expression $A_{\text{coll}} = Cp^2\xi$, where the factor $\xi = 0.7$ takes into account the polarizing effect of the connecting wires (see Methods)¹³. From the data in Fig. 1c, critical coupling is obtained with a period $p = 3.3 \mu\text{m}$, which corresponds to a collection area of $A_{\text{coll}} = 7.5 \mu\text{m}^2$, four times larger than the electrical area, $\sigma = 1.7 \mu\text{m}^2$, of the patch.

The device fabrication process was optimized to generate current solely under the square metallic patches, and not below the

¹Laboratoire Matériaux et Phénomènes Quantiques, Université Paris Diderot, Sorbonne Paris Cité, CNRS-UMS 7162, 75013 Paris, France. ²School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT Leeds, UK. ³ETH Zurich, Institute of Quantum Electronics, Auguste-Piccard-Hof 1, 8093 Zurich, Switzerland.

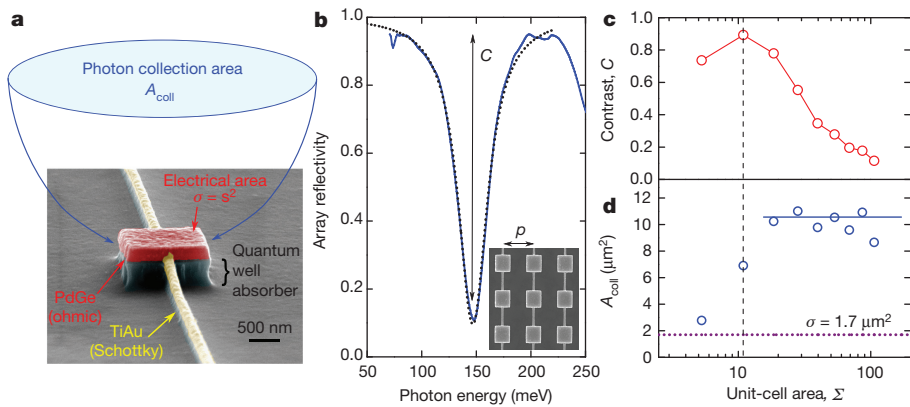


Figure 1 | Device concept. **a**, Double-metal patch antenna, with the various metallic layers employed for electrical contacts (see Methods). The absorbing region contains a 386-nm-thick QWIP structure with five quantum wells doped with Si at a concentration of $n = 7 \times 10^{11} \text{ cm}^{-2}$. In this metamaterial structure, the photon collection area, A_{coll} , is much larger than the electrical area, σ . **b**, Reflectivity spectrum (blue curve) of a patch antenna array with patch size $s = 1.30 \mu\text{m}$ and period $p = 3.30 \mu\text{m}$. The dashed line is a Lorentzian fit used to obtain the absorption contrast C . **c**, **d**, Contrast C (**c**) and collection area A_{coll} (**d**) as a function of the unit-cell area $\Sigma = p^2$ of the array. The observed saturation of A_{coll} is in agreement with theoretical predictions¹³.

150-nm-wide leads connecting them. To this end, we realized ohmic contacts between the patches and the semiconductor layers underneath them using an annealed PdGeTiAu alloy, whereas a Schottky barrier, made by depositing TiAu on the semiconductor surface, prevents the generation of vertical currents between the metallic wire and the semiconductor. Moreover, all cavities were connected with an external wire-bonding pad insulated by an 800-nm-thick Si_3N_4 layer (see Methods). Thanks to all these precautions, the conductive area was reduced to the sum of the areas of the patch resonators, preventing the flow of additional dark current across the device.

To quantify the detector performance, we compared the detector array with a reference device, called ‘mesa’, with the same absorbing region processed into a 200- μm -diameter circle and with light coupling through a 45° polished substrate edge⁷. This comparison revealed the intrinsic photoresponse of the detector (see Methods). In Fig. 2a we compare the peak responsivities for the two configurations, obtained with a calibrated blackbody source at 1,000 °C (see Methods). The mesa device could be characterized only up to 150 K because the photocurrent becomes undetectable at higher temperatures. The array detectors showed a sevenfold enhancement of the responsivity at low temperatures compared with the mesa device and could be characterized up to room temperature, where their responsivity (0.2 A W^{-1}) was comparable with the best responsivity of the mesa device, measured at around 50 K. We were thus able to record photocurrent spectra up to room temperature (Fig. 2b) which was, to our knowledge, the first such measurement with a QWIP operating in the 9- μm -wavelength band using a thermal source.

By quantifying the number of photons absorbed in each configuration (see Methods), we were able to extract the photoconductive gain, g ,

for each structure (Fig. 2c). This gain represents the number of electrons circulating per photon absorbed in the quantum wells^{7,28} and is an intrinsic property of the absorbing region. All our devices (patch antennae nearly resonant with the detector absorption and the mesa device; see Methods) show the same gain as a function of temperature, irrespective of their fabrication geometry, which proves that the material properties are identical for all structures. The photoconductive gain is proportional to the electron drift velocity in AlGaAs barriers⁷ and its temperature dependence is linked to microscopic scattering processes in polar materials^{14,15}. Our results agree with the temperature dependence of the drift velocity described in ref. 14. The derived low-temperature value of the drift velocity is about $6 \times 10^6 \text{ cm s}^{-1}$, as expected in an electric field of 20 kV cm^{-1} for an Al concentration of 20%–30%²⁹. These results account for the drop in the responsivity with increasing temperature observed in Fig. 2a. Above 200 K, the gain g has an almost constant value of 0.2–0.25, of the order of $1/N_{\text{qw}}$. This implies that photoexcited electrons can travel only from one well to an adjacent one, as the mean free path of the electrons is now shorter than the distance between two wells. Interestingly, in this transport regime, a detector based on a single quantum well would be advantageous at high temperatures. These results illustrate how our device gives access to the high-temperature physics of quantum detectors, a regime unexplored so far.

The detector performance is best evaluated through the specific detectivity⁷ $D^* = R\sqrt{A_{\text{det}}}/\sqrt{4egI}$, where R is the responsivity, A_{det} is the area of the detector, e is the electron charge and I is the photocurrent, which is plotted in Fig. 3a for the mesa reference and for the patch devices. The experimental results are compared with those of our model, which describes the influence of the photonic design on the specific detectivity as a function of the temperature¹³. In Fig. 3b we

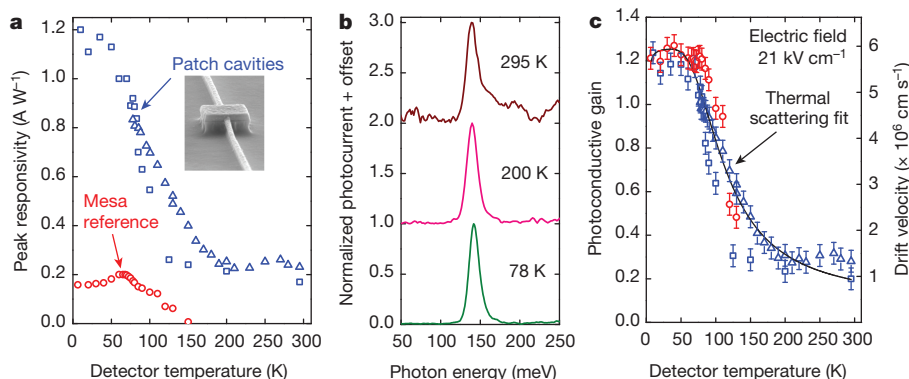


Figure 2 | Detector characterization. **a**, Peak responsivity of QWIP devices, measured with a calibrated 1,000 °C blackbody source. The devices were fabricated in the geometries of the 200- μm -diameter mesa reference device (circles) and resonator arrays with patch sizes $s = 1.35 \mu\text{m}$ (squares) and $s = 1.30 \mu\text{m}$ (triangles). **b**, Photocurrent spectra of the 1.30- μm -patch array at 78 K, 200 K and 295 K, normalized to unity.

A constant offset of unity has been applied for clarity. **c**, Photoconductive gain and electronic drift velocity of the three devices presented in **a** as a function of temperature, for 0.5 V bias voltage (electric field, 21 kV cm^{-1}). The drift velocity is obtained using a quantum-well capture time of 5 ps (see ref. 7 and Methods). The data–symbol correspondence is the same as in **a**. Error bars represent a systematic error from our calibration set-up.

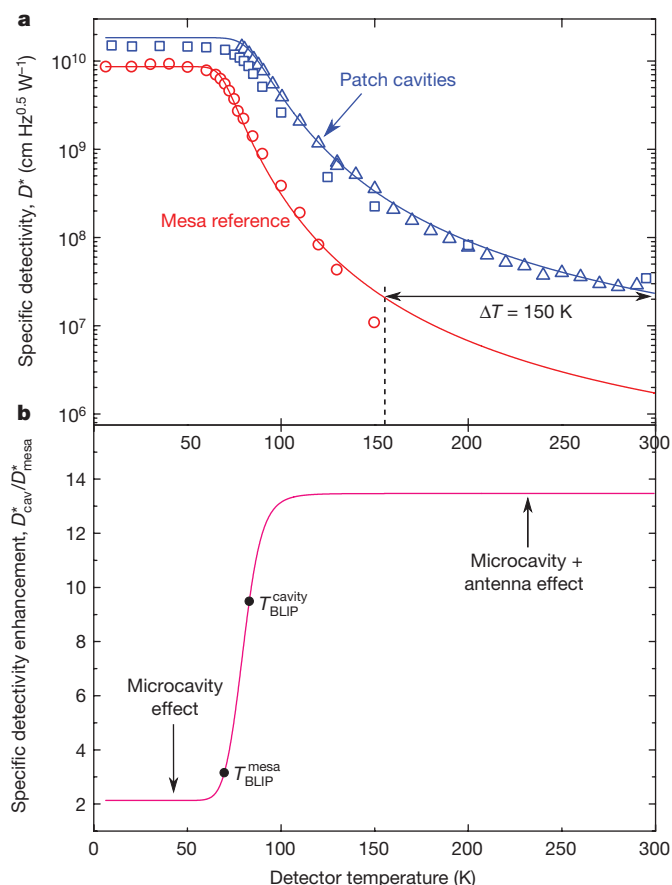


Figure 3 | Specific detectivity as a function of the temperature.

a, Specific detectivity D^* (2π field of view) as a function of the temperature at a bias of 0.5 V for the mesa reference (circles) and two array structures with $s = 1.30\ \mu\text{m}$ (triangles) and $s = 1.35\ \mu\text{m}$ (squares). The red line is a fit of the reference data using $D^*(T) = d_0/[1 + d_1 T \exp(-E_{\text{act}}/k_B T)]^{1/2}$, where T is the temperature, d_0 and d_1 are fitting parameters, $E_{\text{act}} = 120\ \text{meV}$ is the activation energy and k_B is the Boltzmann constant. The blue curve represents the model of quantum detectors embedded in patch resonators described in ref. 13. **b**, Ratio of the specific detectivities of the cavity array with $s = 1.3\ \mu\text{m}$ and the reference mesa. Dots show the corresponding background-limited infrared performance BLIP temperatures: $T_{\text{BLIP}}^{\text{mesa}} = 70\ \text{K}$ (mesa) and $T_{\text{BLIP}}^{\text{cavity}} = 83\ \text{K}$ (patch cavity arrays).

provide the ratio between the specific detectivities of the two configurations. At low temperature, we observe an enhancement by a factor of only two. Here, the dark current is negligible and the main source of noise is the background photocurrent induced by the 300 K blackbody of the environment. In this regime, higher responsivity means also higher background noise, and the specific detectivity enhancement scales with the square root of the responsivity ratio $(R_{\text{array}}/R_{\text{mesa}})^{1/2} = 2.6$. The situation is totally different at high temperature, where the dark current is the dominant contribution to the noise. In this case, the specific detectivity enhancement is:

$$(R_{\text{array}}/R_{\text{mesa}})(A_{\text{coll}}/\sigma)^{1/2} \approx 14 \quad (1)$$

and the performance of the arrays at 300 K is equivalent to that of the mesa reference at 150 K. This doubling of the operation temperature is a considerable improvement, well beyond that predicted from the low-temperature performance of the device. This is due to both the enhancement of the responsivity and the strong suppression of the dark current owing to the antenna effect, expressed by the factor $(A_{\text{coll}}/\sigma)^{1/2}$, because the combination of the microcavity and the antenna effect slows down the decrease of the specific detectivity with temperature, pushing the detector operation to much higher temperatures

than expected¹³. By employing these concepts, we achieved high-temperature operation with relatively high sensitivities.

Owing to its inherent high-frequency response and its reduced electrical capacitance, our device can be used as a heterodyne receiver. In this case, by increasing the power of the local oscillator, we can achieve the highest heterodyne sensitivity limited only by the detector absorption coefficient. This realization is depicted in Fig. 4a, where we show a schematic of the heterodyne arrangement that we used to probe our detector at room temperature. The setup consists of two single-mode distributed feedback (DFB) quantum-cascade lasers (QCLs)¹⁶ operating at $\lambda = 8.36\ \mu\text{m}$. The lasers, used as the signal and the local oscillator, are made collinear by a beam splitter, before impinging on the detector. The latter is wire-bonded with a high-frequency coaxial cable that is connected with a spectrum analyser. Each laser has a linewidth of the order of one megahertz when the current and temperature are stabilized. By adjusting the temperature of each laser, their frequencies are tuned to within a few gigahertz (see Methods).

When the detector is illuminated by both lasers, a clear heterodyne signal appears on the spectrum analyser. In Fig. 4b we show a measurement at 1.06 GHz, with a signal-to-noise ratio of 40 dB. We measured heterodyne signals up to 4.2 GHz, as illustrated in Fig. 4c. Our bandwidth was limited by a strong impedance mismatch between the detector and the external circuit. In Fig. 4d we show the sensitivity of the heterodyne receiver at room temperature. The blue dots correspond to the direct-current (d.c.) saturation curve for the local oscillator, while the red symbols represent the heterodyne signal at 1 GHz as a function of the signal power. The solid blue line is a linear fit for the local oscillator saturation curve. The results show that the detector responds linearly up to 78 mW (about $3.1\ \text{kW cm}^{-2}$) of incident power. Moreover, the linear fit intercepts the 1-Hz integration band at a power of about 0.5 nW, in good agreement with the measured room-temperature specific detectivity shown in Fig. 3a. As can be observed from Fig. 4d, the heterodyne data are well fitted with a line describing a square-root dependence on the power (dashed red line) and can reach a signal-to-noise ratio of unity for an incident power of a few picowatts and an integration time of the order of 10 ms. These results indicate that the heterodyne technique could achieve a sensitivity at $\lambda = 9\ \mu\text{m}$, which is unreachable by any other method at room temperature. We note that in our experiment the photocurrent induced by the local oscillator, $I_{\text{LO}} \approx 0.5\ \text{mA}$, is still dominated by the detector dark current, $I_{\text{dark}} \approx 3.5\ \text{mA}$. By increasing the power of the local oscillator or decreasing the temperature of the detector by a few tens of degrees using thermo-cooled elements, these detectors could reach the heterodyne detection limit, which is defined by their absorption efficiency^{7,13} and the relative intensity noise of the local oscillator³⁰.

We have demonstrated metamaterial photonic detectors operating at room temperature with high sensitivity in the 8–12 μm infrared atmospheric window. Although our detectors show lower d.c. specific detectivity than microbolometers, they have an extremely fast frequency response of tens of gigahertz. In addition, when installed on Peltier elements, the d.c. specific detectivity of our devices is comparable with that of uncooled microbolometers. Using a QCL as a local oscillator, we implemented a heterodyne detection setup and confirmed that these uncooled detectors can operate as coherent heterodyne receivers up to 4.2 GHz. The heterodyne scheme has tremendous potential for sensitive detection in the mid- to far- infrared and could outperform all other competing technologies.

Our devices could be implemented for the detection of coherent signals (lasers), in particular, for free-space high-data-rate transfer¹⁷ and dual-comb spectroscopy³¹, which is an emerging high-resolution spectroscopic technique that requires high-speed detectors. In addition, well established applications, such as optical free-space communications, thermal imaging and environmental remote sensing, will greatly benefit from our coherent sensitive detection. Moreover,

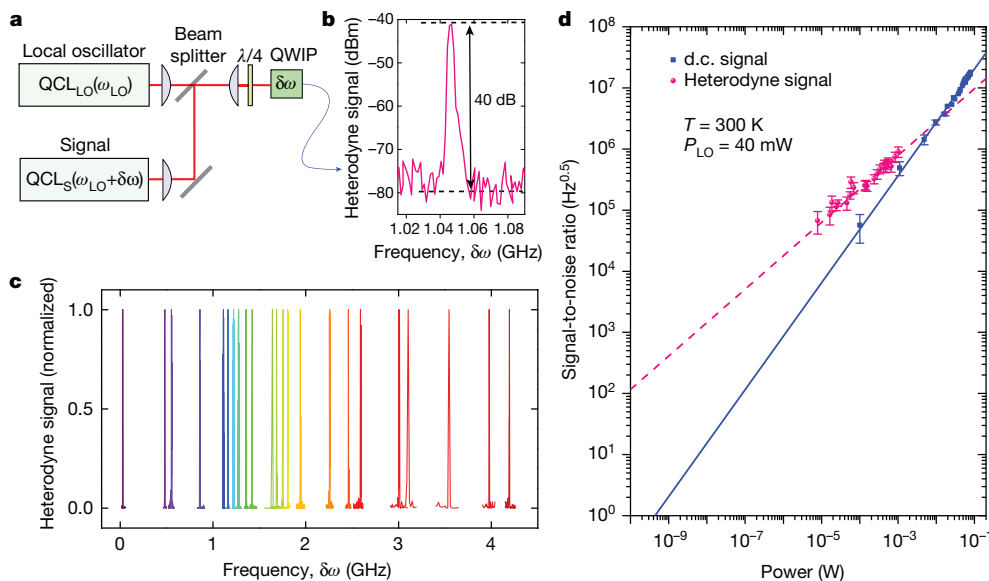


Figure 4 | Tunable heterodyne experiment and results. **a**, Heterodyne arrangement used to probe a cavity-array QWIP at room temperature. The setup contains two single-mode DFB QCLs operating as the local oscillator (ω_{LO}) and the signal ($\omega_{LO} + \delta\omega$), which are made collinear by the beam splitter. The blue arrow indicates electrical connection to a spectrum analyser. **b**, A 40-dB heterodyne power spectrum acquired using a spectrum analyser with 1 MHz resolution bandwidth. **c**, Power spectrum of the heterodyne signal (in the linear scale), normalized to unity. **d**, Logarithmic plot of the signal-to-noise ratio as a function of the power of the signal QCL, for a local oscillator power of 40 mW. The noise of the QWIP is calculated using the measured gain and dark current at room temperature. Each point is an average of ten measurements, and the error bars correspond to the standard deviation.

our heterodyne scheme could also be used for the generation and synthesis of microwaves (up to a few hundreds of gigahertz) with good efficiency, of the order of a few per cent. Finally, these coherent detectors are ideal for implementation in photonic integrated circuits in which a local oscillator is combined with a heterodyne receiver.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 September 2017; accepted 16 January 2018.

Published online 26 March 2018.

- Wood, R. A. in *Infrared Detectors and Emitters: Materials and Devices*, Vol. 8 (eds Capper, P. & Elliott, C. T.) 149–175 (Springer, 2001).
- Rogalski, A. Infrared detectors: status and trends. *Prog. Quantum Electron.* **27**, 59–210 (2003).
- Gunapala, S. D. & Bandara, S. V. in *Semiconductors and Semimetals*, Vol. 62 (eds Liu, H. C. & Capasso, F.) 197–282 (Academic, 2000).
- Mizaikoff, B. Mid-IR fiber-optic sensors. *Am. Chem. Soc.* **75** 258A–267A (2003).
- Martini, R. & Whittaker, E. A. Quantum cascade laser-based free space optical communications. *J. Opt. Fiber Commun. Res.* **4**, 279–292 (2005).
- Henini, M. & Razeghi, M. (eds) *Handbook of Infrared Detection Technologies* (Elsevier, 2002).
- Schneider, H. & Liu, H. C. in *Quantum Well Infrared Photodetectors: Physics and Applications* 72–75 (Springer, 2007).
- Levine, B. F., Choi, K. K., Bethea, C. G., Walker, J. & Malik, R. J. New 10 μm infrared detector using intersubband absorption in resonant tunneling GaAlAs superlattices. *Appl. Phys. Lett.* **50**, 1092–1094 (1987).
- Todorov, Y. et al. Optical properties of metal-dielectric-metal microcavities in the THz frequency range. *Opt. Express* **18**, 13886–13907 (2010).
- Feuillet-Palma, C., Todorov, Y., Vasanelli, A. & Sirtori, C. Strong near field enhancement in THz nano-antenna arrays. *Sci. Rep.* **3**, 1361 (2013).
- Nga Chen, Y. et al. Antenna-coupled microcavities for enhanced infrared photo-detection. *Appl. Phys. Lett.* **104**, 031113 (2014).
- Palaferrri, D. et al. Patch antenna terahertz photodetectors. *Appl. Phys. Lett.* **106**, 161102 (2015).
- Palaferrri, D. et al. Ultra-subwavelength resonators for high temperature high performance quantum detectors. *New J. Phys.* **18**, 113016 (2016).
- Grundmann, M. in *Physics of Semiconductors: An Introduction Including Nanophysics and Applications* 191–195 (Springer, 2010).
- Howarth, D. J. & Sondheimer, E. H. The theory of electronic conduction in polar semi-conductors. *Proc. R. Soc. A* **219**, 53–74 (1953).
- Faist, J. et al. Distributed feedback quantum cascade lasers. *Appl. Phys. Lett.* **70**, 2670–2672 (1997).
- Corrigan, P., Martini, R., Whittaker, E. A. & Bethea, C. Quantum cascade lasers and the Kruse model in free space optical communication. *Opt. Express* **17**, 4355–4359 (2009).
- Argence, B. et al. Quantum cascade laser frequency stabilization at the sub-Hz level. *Nat. Photon.* **9**, 456–460 (2015).
- Vodopyanov, K. L., Chazapis, V., Phillips, C. C., Sung, B. & Harris, J. S. Jr. Intersubband absorption saturation study of narrow III–V multiple quantum wells in the spectral range. *Semicond. Sci. Technol.* **12**, 708–714 (1997).

- Theocharous, E., Ishii, J. & Fox, N. P. A comparison of the performance of a photovoltaic HgCdTe detector with that of large area single pixel QWIPs for infrared radiometric applications. *Infrared Phys. Technol.* **46**, 309–322 (2005).
- Stangier, T., Sonnabend, G. & Sornig, M. Compact setup of a tunable heterodyne spectrometer for infrared observations of atmospheric trace-gases. *Remote Sens.* **5**, 3397–3414 (2013).
- Grant, P. D., Dudek, R., Buchanan, M. & Liu, H. C. Room-temperature heterodyne detection up to 110 GHz with a quantum-well infrared photodetector. *IEEE Photonics Technol. Lett.* **18**, 2218–2220 (2006).
- Schneider, H., Schönbein, C., Bihlmann, G., Van Son, P. & Sigg, H. High-speed infrared detection by uncooled photovoltaic quantum well infrared photodetectors. *Appl. Phys. Lett.* **70**, 1602–1604 (1997).
- Graf, M., Hoyer, N., Giovannini, M., Faist, J. & Hofstetter, D. InP-based quantum cascade detectors in the mid-infrared. *Appl. Phys. Lett.* **88**, 241118 (2006).
- Hofstetter, D. et al. Mid-infrared quantum cascade detectors for applications in spectroscopy and pyrometry. *Appl. Phys. B* **100**, 313–320 (2010).
- Hinds, S. et al. Near-room-temperature mid-infrared quantum well photodetector. *Adv. Mater.* **23**, 5536–5539 (2011).
- Piotrowski, J., Galus, W. & Grudzien, M. Near room-temperature IR photo-detectors. *Infrared Phys.* **31**, 1–48 (1991).
- Liu, H. C. Photoconductive gain mechanism of quantum-well intersubband infrared detectors. *Appl. Phys. Lett.* **60**, 1507–1509 (1992).
- Hava, S. & Auslender, M. Velocity-field relation in GaAlAs versus alloy composition. *J. Appl. Phys.* **73**, 7431–7434 (1993).
- Gensty, T., Elsässer, W. & Mann, C. Intensity noise properties of quantum cascade lasers. *Opt. Express* **13**, 2032–2039 (2005).
- Villares, G., Hugi, A., Blaser, S. & Faist, J. Dual-comb spectroscopy based on quantum-cascade-laser frequency combs. *Nat. Commun.* **5**, 5192 (2014).

Acknowledgements We acknowledge financial support from the FP7 ITN NOTEDEV project (grant number 607521), the ERC grant ADEQUATE, the French National Research Agency (ANR-16-CE24-0020 Project “hoUDINI”) and the EPSRC (UK) projects COTS and HYPERTERAHERTZ (EP/J017671/1, EP/P021859/1). E.H.L. and A.G.D. acknowledge support from the Royal Society and the Wolfson Foundation and thank L. Chen for support with device processing.

Author Contributions D.P., Y.T. and C.S. conceived the experiments, designed the QWIP structure, analysed the data and wrote the manuscript. D.P. fabricated the QWIP devices and performed measurements and data analysis, together with A.B. A.M. and D.G. helped with the heterodyne measurements. A.C. calibrated the blackbody for the responsivity measurements and helped with the characterization of the mesa device. A.V. helped with data analysis. L.L., A.G.D. and E.H.L. grew the QWIP structure and provided the wafer-bonding for the double-metal processing. F.K., M.B. and J.F. provided the DFB QCLs for the heterodyne experiment. All the work was supervised by C.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to C.S. (carlo.sirtori@univ-paris-diderot.fr).

Reviewer Information Nature thanks H. Schneider and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

QWIP fabrication. The QWIP structure is grown by molecular beam epitaxy. It consists of five GaAs quantum wells, each with a thickness of $L_{\text{qw}} = 5.2$ nm and n-doped across the central 4-nm region with Si at a density of $N_d = 1.75 \times 10^{18} \text{ cm}^{-3}$, providing a sheet density of $n = 7 \times 10^{11} \text{ cm}^{-2}$. The quantum wells are separated by $\text{Al}_{0.5}\text{Ga}_{0.5}\text{As}$ barriers of thickness $L_b = 35$ nm. At the top and bottom of this periodic structure, GaAs contact layers are grown, with thicknesses $L_{\text{c,top}} = 100.0$ nm and $L_{\text{c,bottom}} = 50.0$ nm and doping densities $N_{\text{d,top}} = 4.0 \times 10^{18} \text{ cm}^{-3}$ and $N_{\text{d,bottom}} = 3.0 \times 10^{18} \text{ cm}^{-3}$, respectively. The double-metal structures are obtained by wafer-bonding on a GaAs host substrate using 500-nm-thick gold layers and by selectively etching down to an etch-stop $\text{Al}_{0.5}\text{Ga}_{0.5}\text{As}$ layer grown below the bottom contact. As shown in Fig. 1a, the patch antennae are connected by thin, 150-nm metallic wires fabricated using electron-beam lithography (consecutive alignments of lithography steps enable the fabrication of contacts from different metallic alloys). The final structure is obtained by inductively coupled plasma etching of the semiconductor region between the antennae. The Schottky barriers under the thin metallic wires prevent vertical dark current flow between the metal and the semiconductor³². The 45°-facet substrate-coupled geometry consists of a 200- μm -diameter circular mesa, with annealed consecutive layers of Pd, Ge, Ti and Au (deposited by metal evaporation) as a top contact, and annealed Ni, Ge, Au, Ni and Au layers as a diffused bottom contact.

Extended Data Fig. 1 shows a scanning electron microscope image of the quantum detector device built on the basis of our metamaterial photonic concept. The pixel size of the device is $50 \times 50 \mu\text{m}^2$. The external pad is connected with the array by 150-nm-thick wires and is insulated from the bottom ground plane by an 800-nm-thick Si_3N_4 layer. The TiAu pad connects the device with the external circuit by wire bonding.

Reflectivity and photocurrent analysis. Reflectivity and photocurrent spectra were obtained using a Bruker Vertex interferometer. Reflectivity measurements were performed at a 15° incident angle and at room temperature, and the incident light was polarized perpendicular to the 150-nm-thick connecting wires. To obtain the photocurrent spectra, the QWIP devices were mounted in a cryostat with an internal cooled metallic shield and a ZnSe optical window. The photocurrent and responsivity were measured using a blackbody source at 1,000 °C, which was calibrated with an MCT detector. The source was focused onto the detector by two gold parabolic mirrors with $f/1$ and $f/3$ (which denote the ratio between the focal length and the diameter, 5.08 cm, of the parabolic mirrors), providing a typical field of view of 60°. The photocurrent was measured with a lock-in technique using an optical chopper at 1,059 Hz and a shunt resistance connected with the voltage input of a lock-in amplifier (Stanford Research SR1830) without any pre-amplifiers. **Light-polarization dependence.** Our structures support two fundamental modes, TM_{100} and TM_{010} , which are presented in Extended Data Fig. 2a. This figure shows the vertical component E_z of the electric field in the plane of the resonator, as obtained through finite-element simulations. The electric field distribution follows a standing-wave pattern, with a node at the centre of the square and maxima at the edges. The connecting wires perturb the TM_{010} mode slightly, which results in a lower coupling efficiency for this mode. As a result, the total photoresponse of the antenna-coupled device has a co-sinusoidal dependence on the light polarization of a normally incident wave.

In Extended Data Fig. 2b, we plot the peak value of the photocurrent for a structure with $s = 1.30 \mu\text{m}$ as a function of the polarization of a plane wave incident on the array (open circles), with the 90° direction corresponding to the direction of the connecting wires. The angular integral of the cavity photocurrent peak $I_{\text{photo}}(\theta)$ plotted in Extended Data Fig. 2b gives a polarization coupling

coefficient $\xi_{\text{array}} = \int_0^{2\pi} I_{\text{photo}}(\theta) d\theta = 71\%$. The contrast value C of the TM_{100} -

polarized light is obtained from the measurement of Fig. 1b. For comparison, in the same graph we also plot the polarization dependence of the photoresponse measured for the mesa geometry (open squares). Here the 0° direction corresponds to the growth direction of the quantum wells, and the incident wave propagates normally towards the 45° polished facet. This polar plot therefore recovers the inter-subband selection rule, as expected⁷.

Definition of the collection area. Because all incident radiation that is not absorbed is reflected, the contrast C provides the ratio between the incident (P_i) and absorbed (P_a) fluxes for each patch, $C = P_a/P_i$, directly. If Φ_i is the incident photon flux, then the power received by each antenna is $P_i = \Phi_i p^2$ and the power absorbed is $P_a = \Phi_i A_{\text{coll}}$. Then, using the definition of C we obtain $A_{\text{coll}} = Cp^2$. As noted in the main text, we also add a corrective factor of $\xi_{\text{array}} = 0.7$ owing to the polarizing effect of the wires, described in the previous section.

Responsivity, gain and specific detectivity. In Extended Data Fig. 3a we show the responsivity curves as a function of voltage for the mesa and the patch cavity with $s = 1.35 \mu\text{m}$. The decrease of the responsivity with temperature is attributed

to the thermal dependence of the charge-carrier drift velocity and to an increased phonon–electron interaction^{14,15} (see Fig. 2c). The QWIP devices show a typical negative differential photoconductivity, known as the Gunn effect, which involves a decrease of the photocurrent as a function of voltage in electric fields with specific critical intensities, at which inter-valley electron scattering is induced in GaAs⁷.

The responsivity of the mesa can be calculated by considering the voltage (V) dependent photoconductive gain $g(T, V)$ of the active region of the detector and the peak inter-subband energy $E_{21} = 143$ meV, and by including many-body effects:

$$R_{\text{mesa}}(E_{21}, T, V) = \eta_{\text{isb}}(E_{21}) eg(T, V) t_{\text{GaAs}} \xi_{\text{mesa}} / E_{21} \quad (2)$$

where $\eta_{\text{isb}} = 5.0\%$ is the absorption coefficient for the five-quantum-well system in the 45°-facet geometry, $t_{\text{GaAs}} = 0.67$ is the substrate transmission coefficient at 8.6 μm and $\xi_{\text{mesa}} = 0.5$ is the polarization factor (only one polarization of the incident light is coupled with the 45° facet). Analogously to equation (2), we can define¹³:

$$R_{\text{array}}(E_{21}, T, V) = \frac{B_{\text{isb}}(E_{21})}{B_{\text{isb}}(E_{21}) + Q_{\text{ohm}}^{-1} + Q_{\text{rad}}^{-1}} \frac{eg(T, V) C \xi_{\text{array}}}{E_{21}} \quad (3)$$

where $Q_{\text{ohm}} = 4$ and $Q_{\text{rad}} = 22$ represent the ohmic and radiative dissipation of the double-metal cavity, respectively, obtained by reflectivity measurements. Indeed, the Lorentzian fit of the reflectivity resonance shown in Fig. 1b provides its full-width at half-maximum and the sum $Q_{\text{ohm}}^{-1} + Q_{\text{rad}}^{-1}$ and Q_{rad} is calculated from the analytical expression given in ref. 13.

The dimensionless parameter B_{isb} quantifies the energy dissipation through inter-subband absorption and is expressed by a Lorentzian lineshape:

$$B_{\text{isb}}(E) = f_w \frac{E_{\text{pl}}^2}{4E_{21}} \frac{h\Gamma}{(E - E_{21})^2 + (h\Gamma)^2/4} \quad (4)$$

where $f_w = N_{\text{qw}} L_{\text{qw}} / L = 0.067$ is the filling factor of the absorbing quantum wells along their entire thickness L , $E_{\text{pl}} = 47.2$ meV is the inter-subband plasma energy and $\Gamma = 15.0$ meV is the full-width at half-maximum of the photoresponse of the mesa, obtained by a fit to the experimental data. We obtain a similar value $B_{\text{isb}} = 0.07$ for the two resonance cavities with $s = 1.30 \mu\text{m}$ and $s = 1.35 \mu\text{m}$. The absorption coefficient in the antenna-coupled QWIPs is described by the branching ratio $\eta_{\text{array}} = B_{\text{isb}} / (B_{\text{isb}} + Q_{\text{ohm}}^{-1} + Q_{\text{rad}}^{-1}) = 18.9\%$. Using equations (2) and (3) with the measurement data shown in Fig. 2a, we obtain very similar values for the photoconductive gain of the mesa and the array, as shown for the data acquired at 0.5 V (21 kV cm^{-1}) in Fig. 2c. This confirms that the absorbing regions in the two geometries are identical. Furthermore, the data show an exponential decrease of the gain as a function of temperature. According to ref. 7, the photoconductive gain can be defined as:

$$g = \frac{\tau_{\text{capt}} \nu_d}{N_{\text{qw}} L_p} \quad (5)$$

where $\tau_{\text{capt}} = 5$ ps is the capture time, ν_d is the drift velocity, $N_{\text{qw}} = 5$ is the number of quantum wells and $L_p = 40.2$ nm is the length of a period in the structure. The thermal dependence of the gain is related directly to the drift velocity and therefore to the electron mobility. According to ref. 14, we can express the temperature dependence of the photoconductive gain as:

$$g(T) = \frac{1}{\frac{1}{g_0} + \frac{B}{\exp(E_{\text{LO}}/k_B T)} + \left(\frac{E_{\text{ac}}}{k_B T}\right)^{3/2}} \quad (6)$$

Here $E_{\text{LO}} = 36$ meV is the longitudinal optical phonon energy in GaAs, and the fitting parameter $g_0 = 1.25 \pm 0.03$ is the gain at equilibrium (without thermal scattering dependence), with the error obtained from the statistical variance of the fit. The second term in the denominator represents polar optical scattering (see ref. 15), where the parameter $B = 24.4 \pm 1.6$ is a dimensionless polar constant. The third term represents the deformation potential scattering caused by interaction of carriers with acoustic phonons, and the corresponding parameter $E_{\text{ac}} = 0.07 \pm 0.01$ meV characterizes the acoustic deformation potential. Equation (6) fits the experimental data very well, confirming the model.

The photoconductive gain values obtained in this way are used to calculate the specific detectivity as a function of applied voltage at different temperatures, as illustrated in Extended Data Fig. 3b.

Heterodyne measurement. The two beams from the QCLs are made collinear using $f/0.5$ germanium lenses and a beam splitter and then focused onto the detector by an $f/1.5$ lens and a $\lambda/4$ waveplate to avoid optical feedback (Fig. 3a). The two

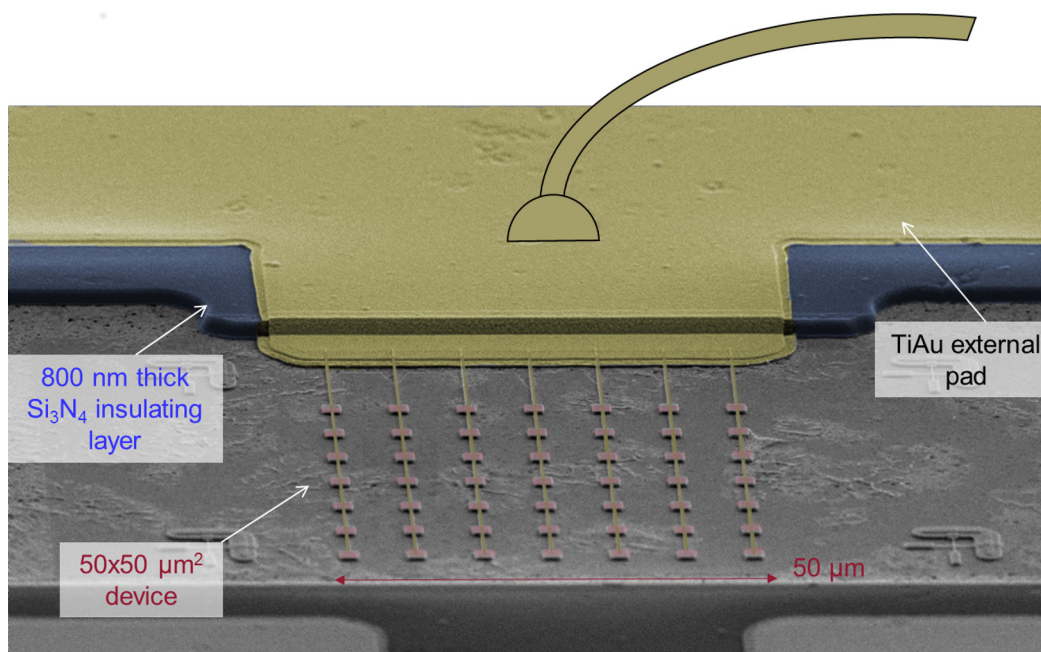
lasers are d.c. biased with a voltage supply and are mounted in two Janis cryostats to stabilize their temperatures using a liquid nitrogen flow. The QWIP is polarized by a Keithley 2450 sourcemeter and the heterodyne signal is sent to a spectrum analyser (Agilent E4407B) using a bias tee. In this arrangement the QWIP detector is at room temperature, without using a cooling system. The QCL used as the local oscillator is kept at a temperature of 254 K and that used for the signal at 293 K. With the temperature stabilized, it is possible to tune the spectral position of the two DFBs by slightly changing the applied d.c. current according to the tuning coefficients $\beta_{LO} = 378 \text{ MHz mA}^{-1}$ and $\beta_S = 413 \text{ MHz mA}^{-1}$ (extracted from a linear fit to the emission frequency of the lasers as a function of temperature and bias).

In the case of a high-power local oscillator, the noise-equivalent power of the heterodyne receiver can be written⁷ as $\text{NEP}_{\text{het}} = E_{21}/(\eta\tau)$, where η is the absorption coefficient of the QWIP and τ is the integration time (defined by the integration bandwidth Δf as $\tau = 1/\Delta f$). For our microcavity-array device we have a theoretical limit of NEP_{het} of less than 1 aW for an integration time of $\tau = 1 \text{ s}$ at 300 K. In the experiment shown in Fig. 4, the signal-to-noise ratio is still mainly limited by the dark current. The square-root fit of the signal-to-noise ratio can be extrapolated to 1, which provides $\text{NEP}_{\text{het}} \approx 10 \text{ fW}$ for an integration time of 1 s ($\text{NEP}_{\text{het}} \approx 1 \text{ pW}$ for an integration time of 10 ms), which is still four orders of magnitude higher than the theoretical limit. These estimations indicate that a high-power local oscillator could achieve sensitivities at the single-photon level at room temperature.

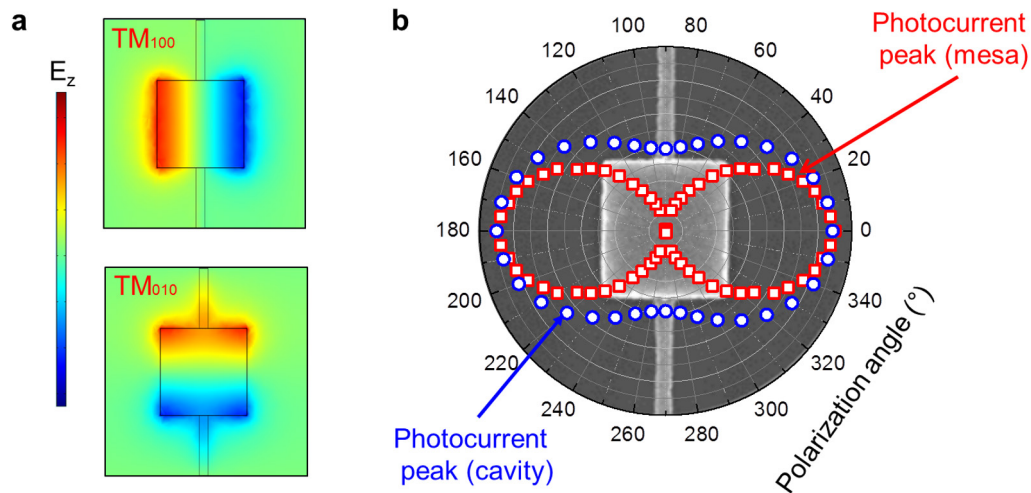
Linearity and heterodyne measurement. In Extended Data Fig. 4 we show the spectra of the two QCLs compared with the room-temperature response of the QWIP with the microcavity-array configuration. The lasers are detuned from the maximum inter-subband absorption, resulting in a detector photoresponse that is half of the maximum achievable. This is an important observation because the responsivity and specific detectivity values that we report in Figs 2 and 3 correspond to the peak values of the detector photoresponse. The background-limited noise-equivalent power is defined as $\text{NEP} = \sqrt{A_{\text{det}}/D^*}$. The detector area A_{det} corresponds to the $50 \times 50 \mu\text{m}^2$ area of the whole array, which is equal to the number of patches N_{patch} multiplied by the unit-cell area $\Sigma = p^2$ of the array. Indeed, in the critical coupling point, all incident radiation is absorbed by the array, and therefore the collection area A_{coll} of each patch coincides with Σ . Using the measured specific detectivity at 295 K for the cavity with $s = 1.30 \mu\text{m}$ at 0.5 V (Fig. 3), we have $D^* = 2.8 \times 10^7 \text{ cm Hz}^{0.5} \text{ W}^{-1}$ and $\text{NEP} = 0.2 \text{ nW Hz}^{-0.5}$. Taking into account the 50% spectral overlap, we obtain $\text{NEP} = 0.4 \text{ nW Hz}^{-0.5}$, which agrees with the value obtained from the linearity measurement shown in Fig. 4d.

Data availability. All data supporting the findings of this study are available within the paper and the Extended Data files.

32. Sze, S. M. & Kwok, Ng. *Physics of Semiconductor Devices* (Wiley, 2011).

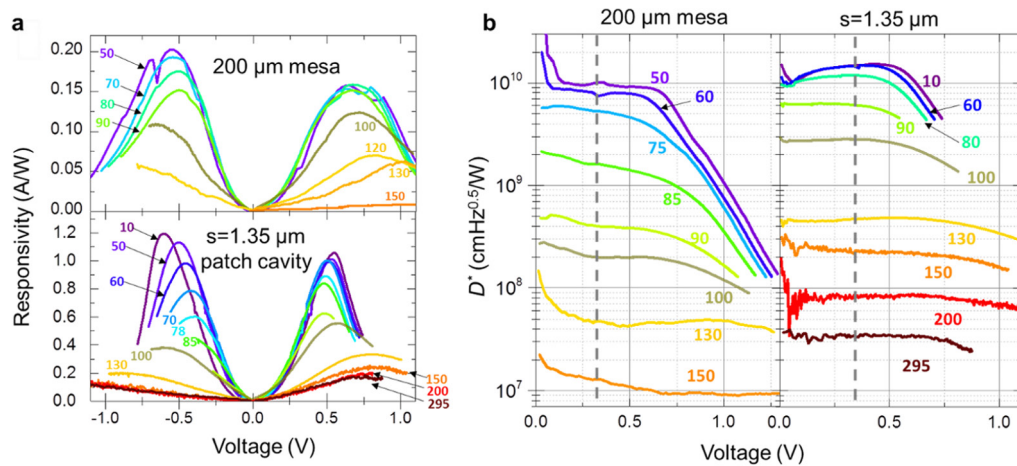


Extended Data Figure 1 | Global view of the device. Scanning electron microscope image of mid-infrared QWIP structure embedded into a $50 \times 50 \mu\text{m}^2$ array of patch resonators. The top TiAu contact is evaporated onto an 800-nm-thick Si_3N_4 insulating layer.



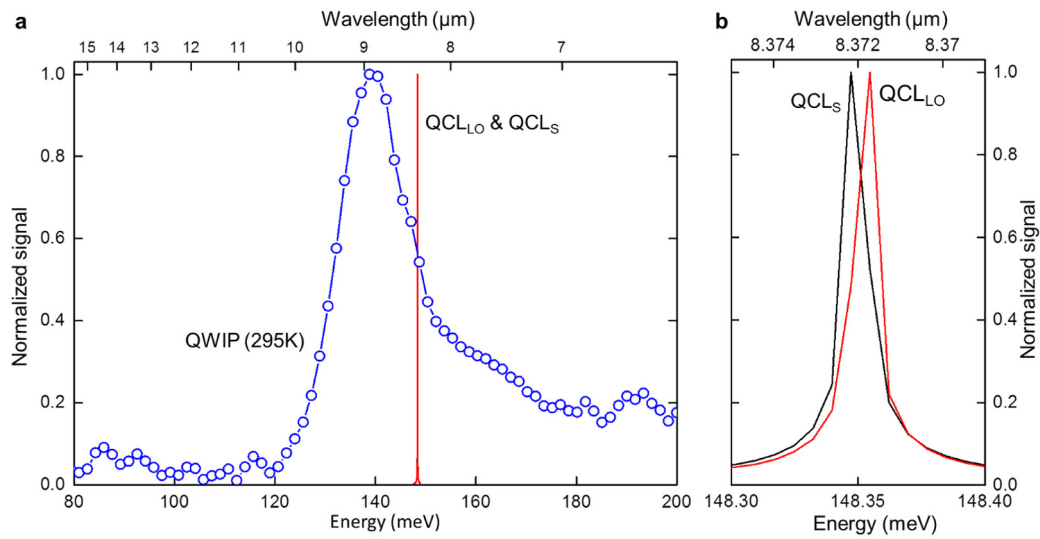
Extended Data Figure 2 | Polarization dependence of the photoresponse. **a**, Finite-element simulation of the E_z component of the electric field coupled with the patch-cavity QWIP for the TM₁₀₀ and the TM₀₁₀ modes. The colour scale represents the amplitude of the electric field and ranges from the absolute maximum (red) to the maximum amplitude in the opposite field direction (blue). **b**, Polar graph of the cavity

photocurrent peak as a function of the wire-grid polarization angle. The photocurrent is normalized by its maximum value, at 0°. The open circles are the results for the cavity array, where the 90° direction corresponds to the connecting wires. The open squares are the results for the mesa geometry, where the 0° direction corresponds to the growth direction of the quantum wells.



Extended Data Figure 3 | Mesa and cavity-array detector characteristics. **a**, Responsivity of the mesa and the $s = 1.35 \mu\text{m}$ antenna-coupled devices as a function of applied voltage. The temperature (in kelvin) of the QWIP

is indicated for each measured curve. **b**, Specific detectivity for the mesa and the microcavity devices as a function of the applied bias at different temperatures.



Extended Data Figure 4 | Spectral characteristics of the two lasers and the QWIP detector. **a**, Emission spectra of the QCLs (QCL_{LO} and QCL_{S}), compared with the room-temperature response of the microcavity QWIP. **b**, Magnification of the spectrum of **a**, showing the two distinct QCL

emission lines. QCL_{LO} was operated at 330 mA and a stable temperature of 293 K, and QCL_{S} was operated at 280 mA and a stable temperature of 254 K.

Molecular nucleation mechanisms and control strategies for crystal polymorph selection

Alexander E. S. Van Driessche¹, Nani Van Gerven^{5,6}, Paul H. H. Bomans^{2,3}, Rick R. M. Joosten^{2,3}, Heiner Friedrich^{2,3}, David Gil-Carton⁴, Nico A. J. M. Sommerdijk^{2,3} & Mike Sleutel^{5,6}

The formation of condensed (compacted) protein phases is associated with a wide range of human disorders, such as eye cataracts¹, amyotrophic lateral sclerosis², sickle cell anaemia³ and Alzheimer's disease⁴. However, condensed protein phases have their uses: as crystals, they are harnessed by structural biologists to elucidate protein structures⁵, or are used as delivery vehicles for pharmaceutical applications⁶. The physiochemical properties of crystals can vary substantially between different forms or structures ('polymorphs') of the same macromolecule, and dictate their usability in a scientific or industrial context. To gain control over an emerging polymorph, one needs a molecular-level understanding of the pathways that lead to the various macroscopic states and of the mechanisms that govern pathway selection. However, it is still not clear how the embryonic seeds of a macromolecular phase are formed, or how these nuclei affect polymorph selection. Here we use time-resolved cryo-transmission electron microscopy to image the nucleation of crystals of the protein glucose isomerase, and to uncover at molecular resolution the nucleation pathways that lead to two crystalline states and one gelled state. We show that polymorph selection takes place at the earliest stages of structure formation and is based on specific building blocks for each space group. Moreover, we demonstrate control over the system by selectively forming desired polymorphs through site-directed mutagenesis, specifically tuning intermolecular bonding or gel seeding. Our results differ from the present picture of protein nucleation^{7–12}, in that we do not identify a metastable dense liquid as the precursor to the crystalline state. Rather, we observe nucleation events that are driven by oriented attachments between subcritical clusters that already exhibit a degree of crystallinity. These insights suggest ways of controlling macromolecular phase transitions, aiding the development of protein-based drug-delivery systems and macromolecular crystallography.

How do protein crystals nucleate? What is or are the pathway(s) from isolated protein molecules to mesoscopic and finally macroscopic crystals? There have been three independent nanometre-scale observations of protein nucleation at solid–liquid interfaces^{13–15}, revealing both direct and indirect pathways, but these works used atomic force microscopy—a surface technique that is blind to events taking place within the liquid. In another approach, *in situ* liquid-cell transmission electron microscopy was used to map the nucleation pathways of calcium carbonate¹⁶ and, more recently, of the protein lysozyme¹⁷, but that technique currently lacks the lateral resolution needed to resolve the structure of the nuclei and the particles that precede them.

To obtain an experimental window onto the formation of a crystal nucleus in liquid at molecular resolution, we use cryo-transmission electron microscopy to image vitrified samples that have been plunge frozen at various time intervals. We study the nucleation pathways of glucose isomerase, a protein with applications in the biofuel and food

industries as a crystalline suspension¹⁸. Depending on the solution conditions, glucose isomerase can crystallize into at least two different space groups¹⁹, or (as we show here) can aggregate into a disordered, gelled state. Using ammonium sulfate as a precipitant, we find that the protein exhibits a polymorph transition from an *I*222 (rhombic) to a *P*₂₁₂₁₂ (prismatic) space group as a function of the precipitant concentration (Extended Data Fig. 1). Turbidity measurements reveal that the induction time for nucleation decreases exponentially as the ammonium sulfate concentration increases from 1.2 M to 1.65 M (Extended Data Fig. 2). However, no conditions are identified that lead to liquid–liquid phase separation or gelation. Cryo-transmission electron microscopy (cryo-TEM) imaging of the earliest quenched sample (plunge frozen after 20 seconds) in 1.5 M ammonium sulfate (a mixed *I*222/*P*₂₁₂₁₂ condition) shows the presence of elongated particle assemblies ('nanorods'; Fig. 1a–e). Given the overall particle dimensions and the electron-microscopy silhouette of the subunits, we identify the building blocks of these nanorods to be single protein molecules (Fig. 1a–c). The nanorods are on average two molecules in width (1.7 ± 1 ; $n = 60$) and 12 molecules in length (12.4 ± 5 ; $n = 60$), with an intermolecular distance of 8.2 ± 0.1 nm ($n = 51$) along the long axis (Fig. 1d and Extended Data Fig. 3). Single-file protein chains are also observed (Fig. 1c), as well as trimers, tetramers and larger polymers at later time points (10–20 min; Fig. 1e). Although successive images show a gradual increase in the nanorod concentration as a function of time (Fig. 1a, b; see Extended Data Fig. 4 for the dependence on ammonium sulfate concentration), there is no increase in their length (Extended Data Fig. 3g).

At around 15 to 30 minutes after protein–precipitant mixing, larger structures begin to emerge. We detect (sub)micrometre-sized fibres of 43 ± 7 nm ($n = 88$) in width (Fig. 1f). The molecular columns that run along the fibre axis have a characteristic centre-to-centre distance of 8.0 ± 0.1 nm ($n = 27$), in line with the spacing measured for the nanorods. The associated two-dimensional fast Fourier transform (2D-FFT) image does not show sharp diffraction spots, but rather diffraction arcs in one or two directions (Fig. 1f, inset). Such arcs indicate that there is local ordering, but also substantial deviation from the crystallographic directions. The aspect ratio of these fibres ranges from 10 to 30, a considerable increase with respect to the aspect ratio of the nanorods (around 6), indicating that fibre broadening is slow compared with elongation. We also see bundles of individual fibres that are making loose lateral contacts with each other (Fig. 1g, h). These bundles have varying degrees of misalignment at the interfibre level, leading to different levels of disorder.

Within the same time frame, faceted nanocrystals start to appear, with morphologies and intermolecular distances that fit the *P*₂₁₂₁₂ and *I*222 space groups (Fig. 2a–e and Extended Data Table 1). The crystallinity of both particle types is reflected in the emergence of sharp diffraction spots in the 2D-FFT. The smallest observed rod-like *P*₂₁₂₁₂

¹Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, IRD, IFSTTAR, ISTerre, F-38000 Grenoble, France. ²Laboratory of Materials and Interface Chemistry and Center of Multiscale Electron Microscopy, Department of Chemical Engineering and Chemistry, Eindhoven University of Technology, PO Box 513, 5600MB Eindhoven, The Netherlands. ³Institute for Complex Molecular Systems, Eindhoven University of Technology, PO Box 513, 5600MB Eindhoven, The Netherlands. ⁴Structural Biology Unit, CIC bioGUNE, Parque Tecnológico de Bizkaia, 48160 Derio, Bizkaia, Spain. ⁵Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. ⁶Structural and Molecular Microbiology, Structural Biology Research Center, VIB, Pleinlaan 2, 1050 Brussels, Belgium.

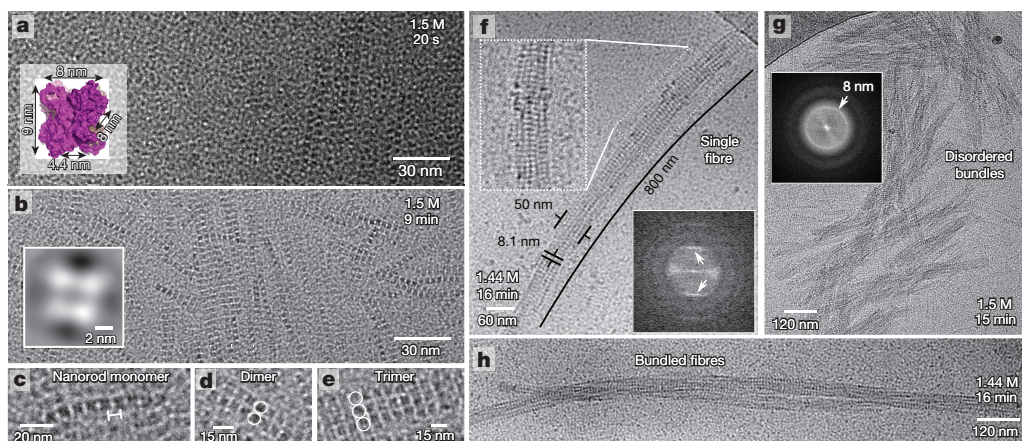


Figure 1 | Pre-nucleation assemblies of glucose isomerase induced by ammonium sulfate. **a, b**, Cryo-TEM image of a glucose isomerase solution in 1.5 M ammonium sulfate, at 20 seconds (**a**) and 9 minutes (**b**) after mixing. The inset in panel **a** shows a van der Waals representation of a glucose isomerase molecule. The inset in panel **b** shows a class-average electron microscopic image of the nanorods' building blocks. **c–e**, Magnified images of a nanorod monomer (**c**), dimer (**d**) and trimer

(**e**) composed of glucose isomerase molecules, with a centre-to-centre distance of 8.2 ± 0.1 nm ($n = 51$, marked in panel **c**). **f–h**, Emergence of larger structures (at 15–30 min). **f**, A micrometre-sized fibre, with the upper inset being a magnified image and the lower inset being the corresponding 2D-FFT image, showing diffraction arcs in two directions. **g, h**, Bundling of fibres into aggregate structures with varying degrees of alignment. The inset in panel **g** is a 2D-FFT image.

crystal measures 380 nm by 120 nm (aspect ratio = 3), and has a width that exceeds those of the fibres (Fig. 2a). The characteristic interplanar spacing parallel to the long crystal axis is 8.1 ± 0.1 nm ($n = 26$), again in line with the value obtained for the nanorods and fibres. The nanorod alignment parallel to the nearest facet of the crystallite in Fig. 2b suggests that oriented attachment is a mode of incorporation into the crystalline phase. For the rhombic *I*222 space group, the smallest crystals that we find have an edge length of ± 100 nm (Fig. 2d, e) and characteristic distances of 5 nm and 7 nm.

With polyethylene glycol (PEG₁₀₀₀ or PEG₁₅₀₀) as the precipitant, glucose isomerase exhibits a similar polymorph transition from rhombic to prismatic crystals, albeit over a relatively narrow PEG concentration range (Extended Data Figs 1, 5). However, the highest PEG concentration produces a different effect to the highest ammonium sulfate concentration, in that glucose isomerase solidifies rapidly into a kinetically arrested gelled state (Extended Data Fig. 6). Cryo-TEM imaging of a range of PEG conditions reveals striking similarities to the

nucleation pathways observed with ammonium sulfate. At 5% (w/v) PEG₁₀₀₀ (an *I*222-only condition), we detect only mesoscale, rhombic crystals that exhibit fringe patterns compliant with the expected interplanar spacing of the *I*222 space group (Fig. 3a and Extended Data Table 1). Under conditions that lead to nucleation of both of the space groups and the gel (86 mg ml^{-1} , 4.5% (w/v) PEG₁₅₀₀; Extended Data Fig. 5b), fibre-like structures appear 2–3 minutes after protein/precipitate mixing; these structures have a characteristic intermolecular distance of 8.0 ± 0.2 nm ($n = 24$) along the long axis, and measure 41 ± 6 nm ($n = 166$) in width (Fig. 3b–e). Interestingly, we observe no nanorods in any of the time points for this sample series, or for any glucose isomerase/PEG sample (Fig. 3b–e). At later time points, we see the grouping of these fibres into structures of increasing dimensions, exhibiting lateral stacking of individual fibres but still separated by a thin solvent layer (Fig. 3d, e). Identical sample replicates that failed to crystallize, but ended up in the gel state, reveal the presence of fibres that are morphologically similar to those described above

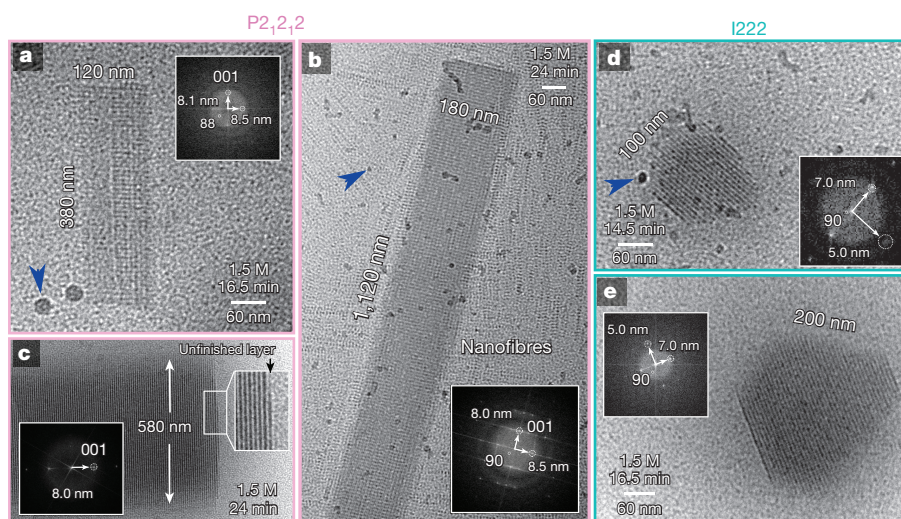


Figure 2 | Emergence of faceted prismatic and rhombic nanocrystals 15–30 minutes after mixing glucose isomerase and ammonium sulfate. **a**, A $P2_12_12$ nanocrystal. Inset, 2D-FFT image, with diffraction spots (measured $8.1 \text{ nm} \times 8.5 \text{ nm}$, 88° ; theoretical (110) plane $7.8 \text{ nm} \times 8.1 \text{ nm}$, 90°). **b**, Larger $P2_12_12$ crystal displaying higher-order diffraction. **c**, Mature $P2_12_12$ crystal with a clear fringe pattern with a spacing of 8 nm). The

magnified image of the facet edge resolves an unfinished molecular layer, suggesting that crystal growth at this point proceeds by incorporation of single glucose isomerase molecules. **d, e**, The smallest detected rhombic, crystalline objects display diffraction spots in two directions, that is, 7 nm and 5 nm at 90° (theoretical (101) plane $6.7 \text{ nm} \times 5 \text{ nm}$; (011) plane $7.3 \text{ nm} \times 4.3 \text{ nm}$). Blue arrows show contamination by ethane.

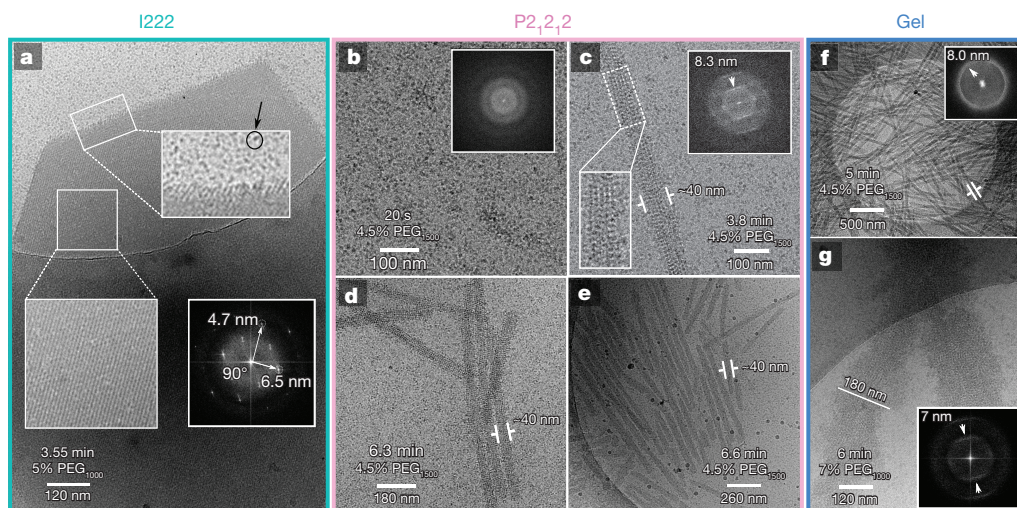


Figure 3 | Cryo-TEM imaging of PEG-induced crystallization of glucose isomerase. **a**, Micrometre-sized rhombic crystals obtained 4 minutes after mixing glucose isomerase and PEG at a low depletion attraction (in 5% PEG₁₀₀₀). Lower left inset, a magnified image of the resolved crystal lattice. Lower right inset, the corresponding 2D-FFT image. Upper inset, magnification of the upper facet edge, resolving free monomers in

solution. **b–e**, Time series of a glucose isomerase/PEG sample with 4.5% PEG₁₅₀₀ reveals a $P2_12_12$ nucleation pathway reminiscent of that observed with ammonium sulfate (Fig. 1), but with a clear absence of nanorods at all measured time points. **f, g**, Formation of a glucose isomerase/PEG hydrogel at an intermediate depletion attraction (**f**, with 4.5% PEG₁₅₀₀) and a high depletion attraction (**g**, with 7% PEG₁₀₀₀).

(of width 44 ± 5 nm; $n = 100$), but at drastically higher concentrations (Fig. 3f). With higher depletion–attraction forces (through the use of 7% PEG₁₀₀₀) but lower protein concentrations (37.5 mg ml^{−1})—conditions that slow down the rate of aggregation—cryoTEM imaging before kinetic arrest (6 min after protein/PEG mixing) reveals the formation of fractal-like aggregates with a distinct lack of rotational order. This can also be seen from the two concentric arcs in a 2D-FFT image (Fig. 3g). The FFT image is reminiscent of those of the disordered fibre bundles observed in the ammonium sulfate experiments, but shows higher packing density, as indicated by the 7.0 ± 0.2 nm ($n = 16$) spacing (Fig. 3g) and broader fibre cross-section (of width 100 ± 50 nm; $n = 20$).

The striking resemblance at the microscopic level between the $P2_12_12$ crystallization pathways (Fig. 1, with ammonium sulfate; or Fig. 3b–e, with PEG) and the gelation pathway strongly suggests that both phases originate from the same precursor states (that is, fibres). This observation prompted us to perform seeding experiments using glucose isomerase/PEG hydrogels, to see whether we could selectively elicit $P2_12_12$ crystals. For this, we transferred a glucose isomerase/PEG hydrogel fragment to a freshly prepared mother liquor solution that leads exclusively to $I222$ crystals (Fig. 5a). Time-lapse imaging of the solution–gel interface reveals the rapid and exclusive formation of $P2_12_12$ crystals on, or protruding from, the gel phase, demonstrating that gels can be used as polymorph-specific seeding agents. If the hydrogel is instead transferred to a similar $I222$ -exclusive condition but a lower PEG concentration (3% (w/v) PEG₁₅₀₀), then the gel phase gradually dissolves as $P2_12_12$ crystals emerge over time (Fig. 5f, g).

Both the early-stage nanorods (formed in high ammonium sulfate concentrations) and the later-stage fibres (formed in high ammonium sulfate or PEG) exhibit high aspect ratios, suggesting that there are substantial differences in lattice-contact strengths ($|C_{ij}|$) for these conditions. To understand the origins of these differences, we analyse the mode of intermolecular bonding within the nanorod structure and compare it with known crystal lattice contacts. Using the glucose isomerase atomic structures for both space groups, we generate nine plausible nanorod models (Extended Data Fig. 7). On the basis of clear discrepancies between the intermolecular distances in these models, and a comparison of the cryoTEM silhouette and the van der Waals model, the only plausible orientation of the nanorod from the $P2_12_12$ space group is in the (001) direction (Fig. 3 and Extended Data Fig. 2).

There are two lattice contact types in the $P2_12_12$ space group— C_1 along the (001) direction, and C_2 along the (110) direction, involving the formation of six and seven hydrogen bonds, respectively (Fig. 4a, b and Extended Data Table 2). The nanorod anisotropy suggests that

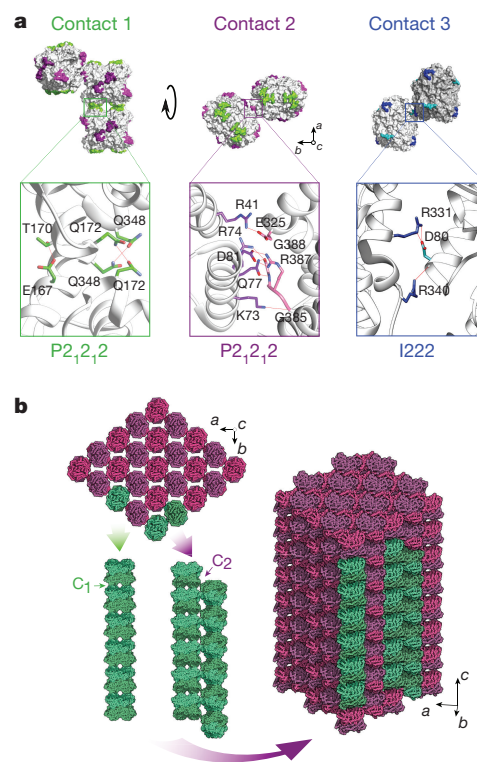


Figure 4 | Lattice contact analysis for both space groups of glucose isomerase. **a**, Contact types for the $P2_12_12$ and $I222$ space groups, with magnifications of the corresponding surface patches and hydrogen bonds, and showing the amino acids at the contact sites (T, threonine; Q, glutamine; E, glutamate; R, arginine; D, aspartate; K, lysine; G, glycine). Hydrogen bonds are shown with solid lines. **b**, Depiction of the molecular columns that run through a $P2_12_12$ glucose isomerase crystal, with the c -axis corresponding to the long axis of the nanorods observed in cryo-TEM. C_1 and C_2 are contacts 1 and 2.

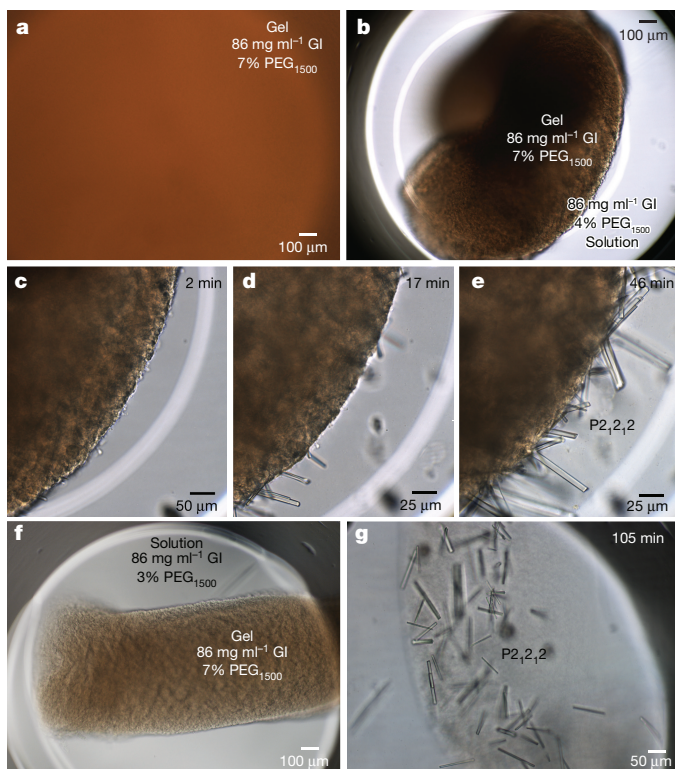


Figure 5 | Gel-mediated nucleation of P_{21212} crystals. **a**, Glucose isomerase/PEG hydrogel formed with 86 mg ml^{-1} glucose isomerase (GI) and 7% (w/v) PEG_{1500} . **b**, An aliquot of the gel shown in panel **a** was transferred to a freshly prepared solution of 86 mg ml^{-1} glucose isomerase and 4% (w/v) PEG_{1500} . **c–e**, Consecutive magnified images of the gel–solution interface, at respectively 2 minutes, 17 minutes and 46 minutes after transfer of the gel aliquot. **f**, An aliquot of the gel shown in panel **a** was transferred to a freshly prepared solution of 86 mg ml^{-1} glucose isomerase and 3% (w/v) PEG_{1500} . **g**, Formation of P_{21212} crystals is linked to gradual dissolution of the gel phase.

$|C_1|$ is much greater than $|C_2|$. This bond hierarchy can be rationalized by considering the salting-out effects induced by ammonium sulfate; these effects include preferential hydration, salt exclusion in the vicinity of the surface²⁰, and increased costs of solvent cavity formation²¹. Sulfate ions are excluded with varying degrees from a macromolecular surface²², with local negative charges contributing strongly to the preferential expulsion²³. This in turn leads to a net attraction when two macromolecules are close to each other, owing to an imbalance in the osmotic pressure²⁴. Thus the strength of the C_1 contact is probably a direct consequence of the 16 negative charges that are buried upon formation, compared with the five such buried negative charges in C_2 . Such symmetry breaking will be less pronounced when the precipitant is PEG, which induces a more isotropic attraction, leading to rhombic nuclei, at low concentrations. On the other hand, PEG-induced depletion attraction is not likely to be perfectly uniform for anisotropic particles, as it will favour protein–protein interactions that maximize the overlap volume²⁵ and, by proxy, the total buried surface area upon complexation. The C_1 contact has the largest difference in accessible surface area (ΔASA), followed by C_2 and C_3 (Extended Data Table 2). We argue that this contributes to the emergence of P_{21212} crystals in intermediate PEG concentrations, where the differences between the contacts become amplified.

An additional level of polymorph control can be gained by means of site-directed mutagenesis: knowing the amino-acid composition of specific lattice contacts allows one to tune their strength. We designed three classes of glucose isomerase mutant that are selectively perturbed in the C_1 , C_2 or C_3 modes of interaction. We predict that mutant proteins with impaired C_1 or C_2 contacts will not form P_{21212} crystals; mutants with

altered C_3 contacts should be $I222$ ‘knockouts’. We used crystallization screening of said mutants to investigate the strategy of polymorph control by mutagenesis. As predicted, mutants with defective C_1 contacts (in these S171W mutants, the amino acid at position 171, serine, was mutated to tryptophan) or defective C_2 contacts (R387A mutants, with arginine 387 mutated to alanine; and GI_His mutants, in which the protein’s carboxy terminus was tagged with a run of histidine residues) no longer produced P_{21212} crystals in the tested conditions, but still nucleated into the $I222$ space group (Table 1 and Extended Data Fig. 8a). The opposite was true of C_3 mutants (R331A plus R340D, where D is aspartate). Seeding experiments using wild-type glucose isomerase microcrystals complement the results of the nucleation trials: wild-type P_{21212} crystals exhibited no growth when transferred to solutions containing C_1 or C_2 mutants; similarly, wild-type $I222$ crystals exhibited no growth in solutions of the C_3 mutant (Table 1). Notably, cryoTEM images of S171W, R387A and GI_His crystals in high concentrations of ammonium sulfate reveal the presence of amorphous aggregates, rather than the nanorod or fibre assemblies seen with the wild-type protein (Extended Data Fig. 8b).

To summarize, when in the presence of a high concentration of ammonium sulfate, the glucose isomerase solution undergoes a rapid decomposition into nanorods that have a quaternary structure similar to the molecular arrangement along the c -axis of the P_{21212} space group. At later time points, fibres are formed (at either high ammonium sulfate or intermediate PEG concentrations) that again have identical intermolecular distances to each other along their long axis. Such high-aspect-ratio assemblies are not observed under conditions that exclusively yield the $I222$ polymorph, nor do they have a structure that is compatible with the crystal lattice of the latter. The fibres are therefore exclusively precursors to the prismatic P_{21212} polymorph. For the ammonium sulfate pathway, our observations suggest that nanorods are the primary building blocks of a next-level self-assembly process that leads to the formation of nanorod oligomers, and subsequently to fibre-like assemblies. Having said that, the data obtained with intermediate PEG concentrations show that fibres can also be formed in the absence of a nanorod phase. Fibres increase in width by lateral attachment, which involves the formation of a large number of interprotein bonds—a complex process that can lead to kinetic traps, as shown by the disorder seen in many fibres (Fig. 1f). Thus, assembly size and crystallinity are order parameters that can evolve independently of each other. We hypothesize that local relaxation from a strained, more disordered state—as seen in many fibrous assemblies—into the crystalline arrangement is associated with an activation barrier that is prohibitively large, yielding disordered fibre assemblies that represent a metastable trap in the protein-assembly pathway. Samples in low PEG concentrations show a total absence of nanorods, higher-order assemblies thereof, or any disordered, liquid-like phases. We detect only faceted crystalline objects, suggesting that $I222$ crystals follow a direct nucleation pathway with monomers as their principal building blocks.

The various pathways seen during the crystallization of glucose isomerase reveal a mechanism of protein polymorph selection that takes place at the earliest measurable stages (20 seconds) of self-assembly (Fig. 6). The primary multimers that are formed have an

Table 1 | Nucleation trials of glucose isomerase mutants and crystal growth tests using wild-type seeds

Mutant	Altered contact	$I222$ nucleation*	P_{21212} nucleation†	$I222$ seed growth*	P_{21212} seed growth†
S171W	C_1	✓	×	—	×
R387A	C_2	✓	×	—	×
GI_His	C_2	✓	×	×	×
R331A/R340D	C_3	×	✓	—	—

*50 mM HEPES 7.0 buffer, 100 mM MgCl_2 , 4% (w/v) PEG_{1000} .

†50 mM HEPES 7.0 buffer, 100 mM MgCl_2 , 1.55 M ammonium sulfate.

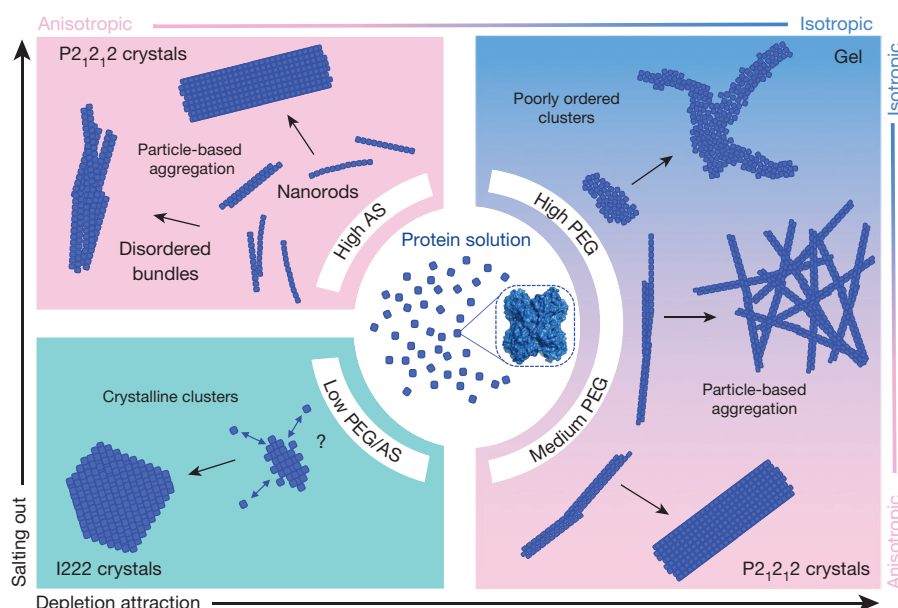


Figure 6 | Proposed model for glucose isomerase crystallization. Top left, high concentrations of ammonium sulfate (AS) induce a strong anisotropic interaction between glucose isomerase molecules, which leads to nanorod formation. These nanorods self-assemble into either disordered fibre bundles or $P2_12_12$ crystals. Bottom left, at low PEG or AS concentrations, glucose isomerase follows a direct nucleation pathway

towards $I222$ crystals, owing to a balance between isotropic repulsion and anisotropic attraction. Bottom right, at medium PEG concentrations, fibres are formed that serve as the precursors to either $P2_12_12$ crystals or gel fibre networks. Top right, at high PEG concentrations, a strong isotropic depletion attraction promotes aggregation into ramified aggregates, resulting in dynamic arrest.

architecture that already resembles the crystalline state. This direct nucleation mechanism can be attributed to the mode of interaction between the glucose isomerase molecules, which is a combination of isotropic repulsion and anisotropic attraction¹³. Such interaction potentials affect the emergent nucleation pathway, as they disfavour disordered dense states^{26,27}. Self-organization of monomers into (pre) critical clusters with a pronounced symmetry determines their subsequent assembly path at their points of inception²⁸. Most unexpectedly, the rod-shaped cluster nucleation pathway for glucose isomerase diverges from the two-step nucleation model for proteins²⁹ that has gained traction recently, but is perhaps more reminiscent of the cluster-cluster interaction at high supersaturation that is described by classical nucleation theory.

To date, control over emerging polymorphs has been based mostly on detailed knowledge of phase diagrams, and has focused predominantly on solubility differences between polymorphs. By contrast, our insights into the mechanism of polymorphism could inspire selection strategies that are geared towards controlling the modes of interaction, including directionality and kinetics. By (de)stabilizing the modes of interaction that are specific to each polymorph, one can control the throughput of the various nucleation pathways, and ultimately influence the yield of the desired polymorph. Such an approach could aid in the development of hydrogel- and crystal-based biotherapeutic agents that require precise control over the outcome of macromolecular phase transitions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 March 2017; accepted 17 January 2018.

1. Siezen, R. J., Fisch, M. R., Slingsby, C. & Benedek, G. B. Opacification of gamma-crystallin solutions from calf lens in relation to cold cataract formation. *Proc. Natl Acad. Sci. USA* **82**, 1701–1705 (1985).
2. Patel, A. *et al.* A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
3. Eaton, W. A. & Hofrichter, J. in *Advances in Protein Chemistry* (eds Anfinsen, C. B., Edsall, J. T., Richards, F. M. & Eisenberg, D. S.) 63–279 (Academic Press, 1990).
4. Ghiso, J. & Frangione, B. Amyloidosis and Alzheimer's disease. *Adv. Drug Deliv. Rev.* **54**, 1539–1551 (2002).

5. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, D301–D303 (2007).
6. Basu, S. K., Govardhan, C. P., Jung, C. W. & Margolin, A. L. Protein crystals for the delivery of biopharmaceuticals. *Expert Opin. Biol. Ther.* **4**, 301–317 (2004).
7. ten Wolde, P. R. & Frenkel, D. Enhancement of protein crystal nucleation by critical density fluctuations. *Science* **277**, 1975–1978 (1997).
8. Pan, W., Galkin, O., Filobelo, L., Nagel, R. L. & Vekilov, P. G. Metastable mesoscopic clusters in solutions of sickle-cell hemoglobin. *Biophys. J.* **92**, 267–277 (2007).
9. Sleutel, M. & Van Driessche, A. E. S. Role of clusters in nonclassical nucleation and growth of protein crystals. *Proc. Natl Acad. Sci. USA* **111**, E546–E553 (2014).
10. Sauter, A. *et al.* Real-time observation of nonclassical protein crystallization kinetics. *J. Am. Chem. Soc.* **137**, 1485–1491 (2015).
11. Sauter, A. *et al.* Nonclassical pathways of protein crystallization in the presence of multivalent metal ions. *Cryst. Growth Des.* **14**, 6357–6366 (2014).
12. Gliko, O. *et al.* Metastable Liquid Clusters in Super- and Undersaturated Protein Solutions. *J. Phys. Chem. B* **111**, 3106–3114 (2007).
13. Sleutel, M., Lutsko, J., Van Driessche, A. E. S., Durán-Olivencia, M. A. & Maes, D. Observing classical nucleation theory at work by monitoring phase transitions with molecular precision. *Nat. Commun.* **5**, 5598 (2014).
14. Chung, S., Shin, S.-H., Bertozzi, C. R. & Yoreo, J. J. D. Self-catalyzed growth of S layers via an amorphous-to-crystalline transition limited by folding kinetics. *Proc. Natl Acad. Sci. USA* **107**, 16536–16541 (2010).
15. Yau, S.-T. & Vekilov, P. G. Quasi-planar nucleus structure in apoferritin crystallization. *Nature* **406**, 494–497 (2000).
16. Nielsen, M. H., Aloni, S. & Yoreo, J. J. D. In situ TEM imaging of CaCO_3 nucleation reveals coexistence of direct and indirect pathways. *Science* **345**, 1158–1162 (2014).
17. Yamazaki, T. *et al.* Two types of amorphous protein particles facilitate crystal nucleation. *Proc. Natl Acad. Sci. USA* **114**, 2154–2159 (2017).
18. Bhosale, S. H., Rao, M. B. & Deshpande, V. V. Molecular and industrial aspects of glucose isomerase. *Microbiol. Rev.* **60**, 280–300 (1996).
19. Gillespie, C. M., Asthagiri, D. & Lenhoff, A. M. Polymorphic protein crystal growth: influence of hydration and ions in glucose isomerase. *Cryst. Growth Des.* **14**, 46–57 (2014).
20. Arakawa, T. & Timasheff, S. N. Preferential interactions of proteins with salts in concentrated solutions. *Biochemistry* **21**, 6545–6552 (1982).
21. Melander, W. & Horváth, C. Salt effect on hydrophobic interactions in precipitation and chromatography of proteins: an interpretation of the lyotropic series. *Arch. Biochem. Biophys.* **183**, 200–215 (1977).
22. Fudo, S., Qi, F., Nukaga, M. & Hoshino, T. Influence of precipitants on molecular arrangement and space group of protein crystals. *Cryst. Growth Des.* **17**, 534–542 (2017).
23. Paterová, J. *et al.* Reversal of the Hofmeister series: specific ion effects on peptides. *J. Phys. Chem. B* **117**, 8150–8158 (2013).
24. Asakura, S. & Oosawa, F. On interaction between two bodies immersed in a solution of macromolecules. *J. Chem. Phys.* **22**, 1255–1256 (1954).

25. Kraft, D. J. *et al.* Surface roughness directed self-assembly of patchy particles into colloidal micelles. *Proc. Natl Acad. Sci. USA* **109**, 10787–10792 (2012).
26. Whitelam, S. Control of pathways and yields of protein crystallization through the interplay of nonspecific and specific attractions. *Phys. Rev. Lett.* **105**, 088102 (2010).
27. Hedges, L. O. & Whitelam, S. Limit of validity of Ostwald's rule of stages in a statistical mechanical model of crystallization. *J. Chem. Phys.* **135**, 164902 (2011).
28. Russo, J. & Tanaka, H. Crystal nucleation as the ordering of multiple order parameters. *J. Chem. Phys.* **145**, 211801 (2016).
29. Vekilov, P. G. Dense liquid precursor for the nucleation of ordered solid phases from solution. *Cryst. Growth Des.* **4**, 671–685 (2004).

Acknowledgements M.S. and N.V.G. acknowledge financial support from the Research Foundation Flanders (FWO) under projects G0H5316N and 1516215N. We thank J. A. Gavira for providing the commercial glucose isomerase sample, S. Van der Verren for assistance with single-particle processing, and H. Remaut for help in designing glucose isomerase mutants.

Author Contributions M.S. and A.E.S.V.D. designed the project and carried out the crystallization and light-scattering experiments. N.V.G. cloned the glucose isomerase mutants and optimized recombinant expression. Mutant proteins were produced and purified by M.S. with the help from N.V.G. Cryogenic freezing and cryoTEM imaging was performed by D.G.-C., P.H.H.B. and R.R.M.J. H.F. advised and co-supervised during cryoTEM imaging. M.S., A.E.S.V.D. and N.A.J.M.S. supervised the study. M.S. and A.E.S.V.D. wrote the paper, with contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.S. (mike.sleutel@vub.be) or A.E.S.V.D. (Alexander.Van-Driessche@univ-grenoble-alpes.fr).

Reviewer Information *Nature* thanks C. Betzel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Protein production and purification. Glucose isomerase was obtained from Hampton Research (from wild-type *Streptomyces rubiginosus*) and received as a crystalline slurry. Small aliquots were dialysed overnight (using Spectra/Por standard regenerated cellulose (RC) tubing, molecular weight cut-off (MWCO) 12–14 kDa; SpectrumLabs) against 10 mM HEPES pH 7.0 buffer plus 1 mM MgCl₂ at 4 °C. The protein solution was concentrated using a centrifugal filter with a MWCO of 100 kDa (Amicon Ultra-15 Cellulose, Milipore) to a typical concentration of 200–250 mg ml⁻¹ and stored at 4 °C. Concentrations were determined by measuring the absorbance at 280 nm using an extinction coefficient ϵ_{280} of 1.074 mg⁻¹ ml cm⁻¹.

Synthetic DNA of full-length wild-type glucose isomerase (UniProt, P24300) with a carboxy-terminal His₆ tag (GI_His), and mutants (S171W, R387A, R331A/R340D) with no carboxy-terminal His₆ tag cloned into plasmid pET22b via *NdeI* and *NcoI* restriction sites, was ordered at GenScript. Recombinant proteins were expressed in *Escherichia coli* strain BL21(DE3) after induction at an optical density (OD)_{600nm} of 0.7 with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) for 3 h at 37 °C. Cells were harvested by centrifugation at 6,238g for 15 min and resuspended in 100 mM Tris-HCl pH 7.3, 1 mM ethylenediaminetetra acetic acid (EDTA, 4 ml per gram of wet cells) supplemented with 5 μ M leupeptin, 1 mM 4-(2-aminoethyl)benzenesulfonyl fluoride (AEBSF), 100 μ g ml⁻¹ lysozyme and 20 μ g ml⁻¹ DNase I, and incubated for 30 min at 4 °C. Subsequently, MgCl₂ was added to a final concentration of 10 mM and cells were lysed by two passages in a Constant System Cell Cracker at 20 kilopounds per square inch (kpsi) at 4 °C; cell debris was removed by centrifugation at 48,400g for 45 min at 4 °C. The cytoplasmic extract was incubated for 10 min at 65 °C and the insoluble fraction was removed by centrifugation at 48,400g for 45 min at 4 °C.

For the non-His-tagged constructs, the supernatant was filtrated through a 0.22 μ m pore filter and loaded on a 5 ml pre-packed Hitrap Q FF column (GE Healthcare) equilibrated with buffer A (50 mM bis-tris-HCl pH 6.0, 10 mM NaCl). The column was then washed with 40 bed volumes of 20% buffer B (50 mM bis-tris-HCl pH 6.0, 1 M NaCl) and bound proteins were eluted with a linear gradient of 20–50% buffer B over 10 bed volumes. Fractions containing wild-type or mutant glucose isomerase—as determined by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE)—were pooled and supplemented with ammonium sulfate to a final concentration of 1.5 M and loaded on a 5 ml pre-packed HiTrap Phenyl HP column (GE Healthcare) equilibrated with buffer A (100 mM Tris pH 7.3, 1.5 M ammonium sulfate). The column was then washed with 40 bed volumes of 25% buffer B (100 mM Tris pH 7.3) and bound proteins were eluted with a linear gradient of 25–85% buffer B over 15 bed volumes. Fractions containing wild-type or mutant glucose isomerase, as determined by SDS–PAGE, were pooled and dialysed (Spectra/Por standard RC tubing, MWCO 12–14 kDa; SpectrumLabs) against 10 mM HEPES pH 7.0 plus 1 mM MgCl₂ overnight at 4 °C (the buffer was replaced twice), and concentrated in a MWCO 100 kDa spin concentrator (Amicon Ultra-15 Cellulose; Milipore) to a typical final concentration of 30 mg ml⁻¹. (For the S171W mutant, ϵ_{280} = 1.198 mg⁻¹ ml cm⁻¹). Cleared cytoplasmic extracts of GI_His were loaded on a 5 ml pre-packed HiTrap Ni-NTA column (GE Healthcare) equilibrated with buffer A (50 mM Tris-HCl pH 7.3, 500 mM NaCl and 20 mM imidazole). The column was then washed with 40 bed volumes of buffer A, and bound proteins were eluted with a linear gradient of 0–75% buffer B (50 mM Tris-HCl pH 7.3, 500 mM NaCl and 500 mM imidazole) over 15 bed volumes. Fractions containing GI_His were pooled, dialysed and concentrated as indicated above.

Glucose isomerase crystallization. For a typical crystallization experiment with wild-type glucose isomerase, the protein stock solution was first diluted to a concentration of 75 mg ml⁻¹ in 50 mM HEPES pH 7.0 and 100 mM MgCl₂, and then mixed at 22 °C with an equal volume of a buffered ammonium sulfate, PEG₁₀₀₀ or PEG₁₅₀₀ solution that was at a concentration twice that desired after mixing. Final concentrations ranged from 0.5 M to 1.75 M ammonium sulfate, and from 3% to 7% (w/v) PEG₁₀₀₀ or PEG₁₅₀₀. Space-group determination was based on the distinct crystal morphologies of both space groups (Extended Data Fig. 1a), using a wide-field optical microscope. The phase diagrams in Extended Data Fig. 1b, c were determined by setting up triplicate crystallization tests using the microbatch-under-oil method (with Nunc 72 microwell minitrays; Sigma Aldrich) at 22 °C, with 10 μ l drops of mother liquor.

Precipitant dependence of glucose isomerase crystallization. We began by mapping out the concentration dependence of glucose isomerase polymorphism for the precipitants used here. Our starting point was 50 mM HEPES pH 7.0 plus 100 mM MgCl₂, which yields exclusively rhombic (I222) crystals for a broad range of protein concentrations (20–75 mg ml⁻¹). By supplementing that condition with ammonium sulfate in 100–200 mM increments, from 0.5 M to an upper limit of 1.75 M final concentration, we saw a gradual shift from rhombic to prismatic (P2₁2₁2)

crystals (Extended Data Fig. 1b). Similarly, by supplementing the base condition with either PEG₁₀₀₀ or PEG₁₅₀₀ to a final concentration of 3% to 8.5% (w/v), we recorded a gradual shift from I222 to P2₁2₁2 crystals; at the higher PEG concentrations, a dense gel phase was formed. We note that there was a narrow PEG concentration range (for PEG₁₅₀₀, from 4.5% to 5.5%) where I222 crystals, P2₁2₁2 crystals and gels were observed simultaneously (see Extended Data Fig. 5 for a detailed microscopic record). Gelation depended only weakly on glucose isomerase concentration: the gelation line occurs as a vertical in Extended Data Fig. 1.

Induction-time measurements. We determined induction times (t_{ind}) for glucose isomerase crystallization as a function of ammonium sulfate concentration (1.2–1.65 M) by following the change in absorbance of freshly prepared supersaturated solutions. We monitored the increase of absorbance in the reacting solutions by *in situ* time-resolved ultraviolet-visible spectroscopy (Agilent Cary 300E spectrophotometer). Measurements were carried out at a wavelength of 500 nm (absorbance of individual glucose isomerase molecules is minimal at this wavelength) and performed in poly (methyl methacrylate) (PMMA) cuvettes located inside a peltier thermostated cell module at 20 °C. The time elapsed between the mixing of protein and salt solutions and the first observed change in turbidity was taken as the induction time. This point was determined through linear fitting of the sigmoidal absorbance curve near the inflection point and determining the intersection with the x-axis.

To obtain a general idea about the nucleation kinetics and to estimate the nucleation induction time, we monitored the turbidity of the crystallizing solution between 1.2 M and 1.65 M ammonium sulfate, and obtained an exponential dependence of t_{ind} on ammonium sulfate concentration (Extended Data Fig. 2). We later used these simple estimates of t_{ind} as a guide for the preparation of cryo-TEM samples, to determine the desired number of samples and time intervals for each condition. We also note that, under high ammonium sulfate concentrations, the system undergoes near-spinodal decomposition with respect to the crystalline phase. Wide-field light microscopy imaging confirms that, even under these conditions, glucose isomerase solidifies exclusively into the (P2₁2₁2) crystalline state. We find no evidence that any amorphous solid states are formed; nor is there any indication that a liquid–liquid phase boundary is crossed.

Time-resolved dynamic light scattering. We collected intensity correlation functions of mixed protein–precipitant solutions at 20 °C, using 10 mm cylindrical cuvettes at an angle of 90°, and employing an ALV-CGS-3 static and dynamic light scattering (DLS) device with a 22 mW helium–neon laser (wavelength 632.8 nm). We collected data in a pseudo cross-correlation set-up in order to minimize the contribution of dead time effects and of photomultiplier-tube after-pulsing to the recorded signal. The intensity autocorrelation function $g_2(\tau) - 1$, with τ being the delay time, is connected to the electric-field correlation function $g_1(\tau) - 1$ through the Siegert relation³⁰:

$$g_2(\tau) = B(1 + \beta |g_1(\tau)|^2)$$

where B is the baseline of the correlation function at infinite delay, and β is the function value at zero delay. For samples containing PEG₁₀₀₀, we used the following double-exponential function (equation (1) is used to fit $g_1(\tau)$ at time points before kinetic arrest, and equation (2) is a stretched exponential used after gelation has occurred):

$$g_1(\tau) = A_1 e^{-\Gamma_1 \tau} + A_2 e^{-\Gamma_2 \tau} \quad (1)$$

$$g_1(\tau) = e^{-\Gamma \tau^p} \quad (2)$$

where p is a fitting parameter, and $\Gamma = Dq^2$ is the decay rate defined by the diffusion coefficient D of the particles and the magnitude of the scattering vector $q = 4\pi n/\lambda \sin(\theta/2)$ at the scattering angle θ .

We collected time-lapse DLS acquisitions to follow, in real-time, the crystallization of filtered glucose isomerase solutions in 50 mM HEPES pH 7.0 plus 100 mM MgCl₂, using a 1.5 M ammonium sulfate solution and 48 mg ml⁻¹ glucose isomerase. The time evolution of the intensity correlation curve is shown in Extended Data Fig. 6a. Processing of the raw curves with the ALV-correlator software (ALV-7004 v3.0.5.1), using the regularization method, yields hydrodynamic radii of the various glucose isomerase populations that form in solution. Thirty seconds after glucose isomerase/ammonium sulfate mixing, light scattered by the glucose isomerase monomers only was collected (measured hydrodynamic radius R_h = 8 nm). Over the course of minutes, a second shoulder started to form in the correlogram, with the earliest measurable species (at 4 min) corresponding to micrometre-sized particles (denoted as ‘clusters’). These species rapidly grew until they completely dominated the recorded signal (by 14 min). Visual inspection at this point showed that the sample had become opaque. *Ex situ* wide-field microscopy analysis after 30 min confirmed the presence of P2₁2₁2 nanocrystals with a

small minority of I222 crystals. On the basis of the typical nanorod dimensions determined by cryoTEM (length 100 nm; aspect ratio 6), we predict—following the corrections described in ref. 31—that they would have an apparent hydrodynamic radius of ± 45 nm. Given the results discussed above, we conclude that we could not detect any light scattered by particles in this size range.

We also tested conditions that do not yield an ordered solid, but instead lead to a kinetically trapped gel state. Time-resolved DLS of a solution of 50 mM HEPES pH 7.0, 100 mM MgCl_2 , 7% (w/v) PEG₁₀₀₀ and 25 mg ml⁻¹ glucose isomerase showed that the intensity auto-correlation function (ACF) could be fit at early time points with a double-exponential decay (with a fast-diffusing population corresponding to monomers, and a slowly diffusing population corresponding to clusters that grow as a function of time). At later stages a stretched exponential was required to reproduce the ACF (Extended Data Fig. 6b, c). Stretched exponentials indicate a hierarchy of fluctuations on all length scales and are a well known characteristic of gels³². Using optical microscopy, we obtained a visual confirmation of the gelled state. The inset of Extended Data Fig. 6c clearly resolves the pores that are present in the mesh of fibres in the kinetically arrested state.

Seeding experiments using glucose isomerase hydrogels. Crystallization trials using PEG as a precipitant showed that glucose isomerase exhibits I222/P2₁2₁2 polymorphism over a narrow concentration range (Extended Data Figs 1, 5). For concentrations lower than or equal to 4% (w/v) PEG₁₅₀₀, we observed only I222 crystals. Conversely, for concentrations higher than or equal to 6% (w/v) PEG₁₅₀₀, we obtained opaque glucose isomerase hydrogels. At 4.5% (w/v) PEG₁₅₀₀, I222/P2₁2₁2 polymorphism occurred, with strongly varying nucleation densities for the P2₁2₁2 space group; in some cases a gelatinous phase also formed that seemed to enter into competition with the crystalline phases. We observed a similar transition regime for PEG₁₀₀₀, but shifted towards higher PEG concentrations (Extended Data Fig. 1). We transferred a small gel fragment grown at 7% (w/v) PEG₁₅₀₀ (Fig. 5a) to a freshly prepared solution that was identical in composition but of a lower PEG₁₅₀₀ concentration ((w/v) 4%; Fig. 5b). Time-lapse imaging of the gel–solution interface revealed the rapid and exclusive formation of P2₁2₁2 crystals on, or protruding from, the gel phase (Fig. 5c–e). Transferring a gel fragment to 3% (w/v) PEG₁₅₀₀, however, led to the gradual dissolution of the gel phase as P2₁2₁2 crystals emerged over time (Fig. 5f, g).

Crystallization of GI_His and mutant proteins. To gain more control over the polymorph selection process, we designed and produced glucose isomerase mutants that we predicted to affect space-group-specific intermolecular contacts while leaving all other contacts unchanged. We had three different types of mutant, impairing C₁ contacts (the S171W mutant, with steric inhibition), C₂ contacts (the R387A mutant, with a salt bridge removed, and GI_His, with steric inhibition) or C₃ contacts (the R331A/R340D mutant, with charge inversion). We predicted that mutants with defective C₁ and/or C₂ interactions would form exclusively I222 crystals, whereas impaired C₃ constructs would form just P2₁2₁2 crystals.

We gauged our ability to control polymorph selection through site-directed mutagenesis by setting up crystallization trials for the new constructs, using conditions that lead (almost) exclusively to either I222 or P2₁2₁2 crystals with wild-type glucose isomerase. Thus, 50 mM HEPES pH 7.0, 100 mM MgCl_2 and 1.55 M ammonium sulfate leads predominantly to P2₁2₁2 crystals, whereas 50 mM HEPES pH 7.0, 100 mM MgCl_2 and 4% (w/v) PEG₁₀₀₀ favours the nucleation of the I222 space group. If no crystallization (of either space group) could be induced with the selected mutant under these conditions, we set up grid screens by varying the precipitant concentration. If only one space group could be obtained after such a screening, we classified the tested mutant as I222-negative or P2₁2₁2-negative (Table 1 and Extended Data Fig. 8a). We note that any crystallization screening is inherently finite, and therefore cannot be used to conclusively rule out the absence of a particular polymorph throughout all of chemical space. Hence, as an auxiliary method, we set up seeded crystallization tests using pre-grown wild-type I222 or P2₁2₁2 glucose isomerase crystals, which we then washed in their corresponding mother liquors to remove any soluble glucose isomerase species, and transferred to an identical mother liquor solution supplemented with 10 mg ml⁻¹ of the respective mutant. We monitored the growth of these seed crystals over time using wide-field microscopy (Table 1).

Cryo-TEM. For cryo-TEM, we used 200-mesh copper grids with Quantifoil R 2/2 holey carbon films (Quantifoil Micro Tools GmbH). We prepared samples using an automated vitrification robot (FEI Vitrobot Mark III) for plunging in liquid ethane³³. Before use, all TEM grids were surface plasma treated for 40 seconds using a Cressington 208 carbon coater. We studied the samples with the Technische Universiteit Eindhoven/FEI cryoTITAN (www.cryotem.nl) operated at 300 kV, equipped with a field emission gun (FEG), a post-column Gatan energy filter (GIF) and a post-GIF 2k × 2k Gatan charge-coupled-device camera. We choose t_0 as the moment at which we induced supersaturation with respect to the crystalline phase (that is, when we mixed the protein with the precipitant solution)

and t_{end} as the time at which crystals became detectable using light microscopy. The exact time point of the samples as indicated in the main text was defined as the moment (after blotting excess liquid) when the electron-microscopy grid was plunged into the liquid ethane. The selected solution conditions represent a compromise between the nucleation density and the overall rate of transformation—that is, for TEM one needs on the one hand a high enough particle density, and on the other slow enough kinetics to manage the cryogenic-quenching at constant time intervals (roughly 2 minutes). We acquired images in low-dose mode at a magnification of either ×24,000 with a nominal defocus of $-5\ \mu\text{m}$, or ×11,500 and $-10\ \mu\text{m}$ defocus.

Single-particle data processing and projection approximation. We determined the defocus of the micrographs by using a script developed in-house (written by R. Efremov). We manually picked and stacked 1,240 particles in E2BOXER. A low-pass Gaussian filter was applied to remove excessive high-frequency noise, and the contrast was inverted before classification. We carried out two-dimensional class averaging by K-means classification using a soft circular mask, and then performed a multi-reference alignment using SPARX³⁴.

To approximate a low-dose cryoTEM projection of the rod assemblies (Extended Data Fig. 8), we used a Protein DataBank (PDB) model. From the PDB model (containing all atom coordinates), we created a three-dimensional density map of the rod via Chimera 1.12 (using the molmap command). Each atom is described as a three-dimensional Gaussian distribution of width proportional to the resolution (3 nm at $-5\ \mu\text{m}$ defocus) and amplitude proportional to the atomic number. The pixel size was set to 0.4 nm, which is close to the pixel size of acquisition (0.38 nm). We summed this three-dimensional intensity map in Matlab along the y -direction perpendicular to the rod length, creating a density projection of the structure. The TEM image was approximated by subtracting the density projection from a flat background image containing Poisson noise (mean intensity = 100 electrons per pixel, as the beam intensity during cryoTEM imaging). Fresnel fringes (white lines surrounding glucose isomerase) arising from the applied underfocus during imaging were not included in the simulation.

Distances along nanorods, fibres, crystals and gel fibres. To obtain estimates of the intermolecular distances within glucose isomerase's nanorod structures, we plotted the power spectrum of the greyscale values along the long axis of a single nanorod, and identified two dominant frequencies that correspond to the characteristic intermolecule and intramolecule distances (Extended Data Fig. 3a–d). In a second approach, we calculated 2D-FFTs using ImageJ 1.50i (ref. 35) from an entire TEM image containing numerous nanorods lying in random orientations (Extended Data Fig. 3a). This orientational averaging yielded an FFT image containing two concentric circles, whose radii again corresponded to the intermolecule and intramolecule distances (Extended Data Fig. 3e). We applied orientational averaging to the 2D-FFT (Extended Data Fig. 3f) and took the inverse frequencies of the two maxima. Applying this second approach to 51 images, we obtained a value of 8.2 ± 0.1 nm for the intermolecular distance along the long nanorod axis (Extended Data Fig. 3g). We used a similar method to determine the characteristic distance within the fibre structures and for the nanocrystals, but used selections corresponding to specific regions of interest (fibre or crystal outline). Also, in the orientational averaging step of the 2D-FFT, we calculated the radial profile over a range of 20° and 5° for fibres and crystals, respectively, instead of 180° for the nanorods. Using 27 fibres, we obtained a characteristic distance of 8.0 ± 0.1 nm along the long axis, and using 29 crystals, we measured 8.1 ± 0.1 nm in the c -direction. For the gel fibres, we integrated over 20° and obtained a spacing of 8.0 ± 0.2 nm (4% (w/v) PEG₁₅₀₀) and 7.0 ± 0.2 nm (7% (w/v) PEG₁₅₀₀), on the basis of 24 and 16 measurements, respectively.

Crystallographic analysis. As starting models for the atomic structures of glucose isomerase in the P2₁2₁2 and I222 space groups, we used the biological assembly models of entries 1OAD and 9XIA in the RCSB PDB (<https://www.rcsb.org>). We generated nearest crystallographic neighbours of the glucose isomerase molecule for both space groups using Chimera 1.11.2rc (Fig. 4). We identified residues that partake in lattice contacts by calculating the accessible surface area (ASA) on a per-residue level, using AREAIMOL of the CCP4 software suite³⁶. We determined ASAs for both of the starting models, and for the models consisting of glucose isomerase and its nearest neighbour, by using a probe radius of 1.4 Å. Residues with a non-zero ΔASA are (partially) buried in the bound complex and therefore considered to be part of the lattice-contact patch. We identified hydrogen-bond pairs with the FindHBond tool in Chimera 1.11.2rc using default settings, and salt bridges using the PDBEPIA (<http://www.ebi.ac.uk/pdbe/pisa/>) and the 2P2I (http://2p2idb.cnrs-mrs.fr/2p2i_inspector.html) protein-interaction web servers.

Given these two models (1OAD and 9XIA) of glucose isomerase for both space groups, we used crystallographic symmetry operations to generate a number of plausible candidates for the experimentally observed nanorods. For this, we

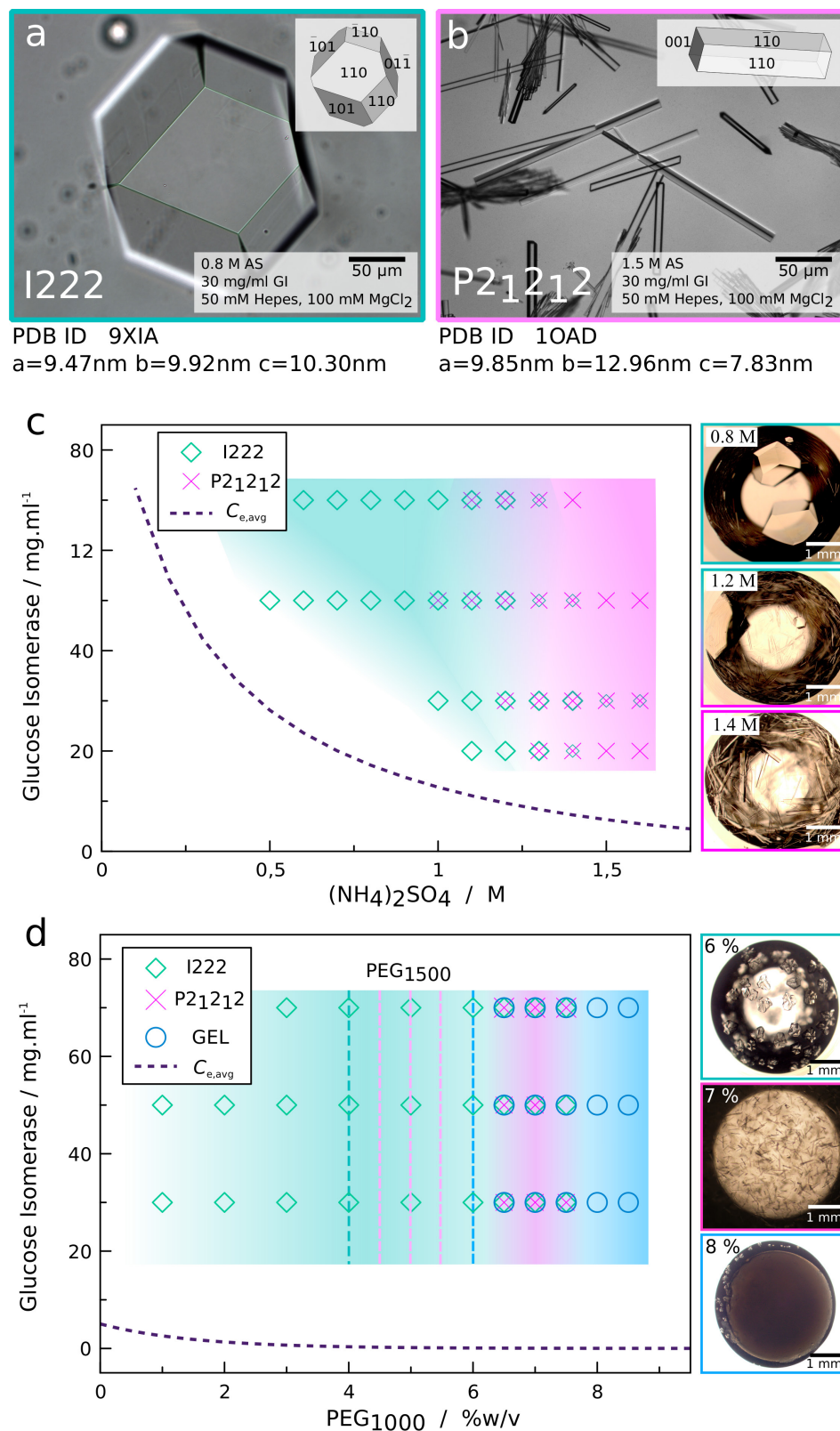
identified the various classes of lattice contact (see below) that exist in both space groups, and applied translational and rotational symmetry operations using the Supercell plugin (<https://pymolwiki.org/index.php/Supercell>) to Pymol to construct linear glucose isomerase sequences in space. A basic requirement that we set is that adjoining glucose isomerase molecules must be in direct contact with each other, be it through van der Waals, hydrogen-bond or electrostatic interactions. We discarded loose packing structures with water-mediated contacts only (Extended Data Fig. 8). Next, we compared the intermolecule distances and particle silhouettes to identify a potential match with the nanorods that were imaged using cryoTEM. Nanorods constructed along the (100), (010), (001) and (011) directions of the $I222$ spacegroup, and along the (100), (010) and (110) directions for the $P2_12_12$ spacegroup, all have a helical ultrastructure that has a pitch defined by the respective unit-cell dimensions. The only linear array of glucose isomerase molecules that could be generated is that in the (001) direction for the $P2_12_12$ space group, and the (111) direction for $I222$. Careful comparison of the orientations of the molecules with respect to the nanorod axis led us to conclude that the $P2_12_12$ (001) model is the most plausible. Indeed, juxtaposing the cryoTEM image of a single nanorod with the van der Waals surface representation of our $P2_12_12$ (001) nanorod model showed good agreement between the two. Also, the $P2_12_12$ (001) nanorod had an intermolecular distance of 7.83 nm when using PDB 1OAD as a reference structure. The nanocrystals that we grew, however, were less compact. The intermolecular distance (along the c -axis; based on the fringe pattern in the cryoTEM images) is 8.1 ± 0.1 nm—a good match to the experimental intermolecular distance within the nanorods (8.1 ± 0.2 nm). We therefore conclude that the nanorods that are formed in high ammonium sulfate concentrations are linear molecular arrays that can also be found along the c -axis of a mature $P2_12_12$ glucose isomerase crystal.

Lattice contacts. For the $P2_12_12$ space group, we identified two types of lattice contact that involve three different surface patches, designated P1, P2a and P2b (Extended Data Table 2). Contact 1 (C_1) is made in the (001) direction by the self-recognition of patch P1, and is duplicated owing to the non-crystallographic

twofold symmetry of the glucose isomerase tetramer. The total contact therefore includes the formation of 2×3 hydrogen bonds and has a ΔASA of 844 \AA^2 . Contact 2 (C_2) along the (110) direction is formed by the binding of P2a with P2b, involving the formation of seven hydrogen bonds and two salt bridges, and encompassing a total ΔASA of 622 \AA^2 . For the $I222$ space group, there is just one lattice-contact type (contact 3, C_3), which involves two surface patches, 1a and 1b, making three hydrogen bonds and resulting in a net ΔASA of 372 \AA^2 . Note that these patches are unique to their respective space groups, although P2a and 1a share one amino acid (D81).

Data availability. We declare that the data supporting the findings of this study are available within the paper and the Extended Data figures and tables. Further data are available from the corresponding authors upon request.

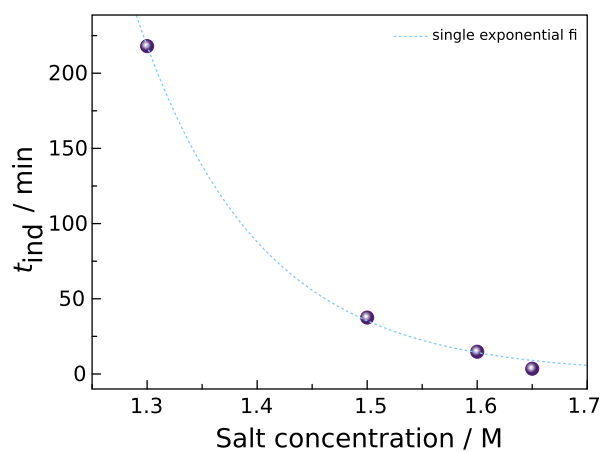
30. Berne, B. J. & Pecora, R. *Dynamic Light Scattering: With Applications to Chemistry, Biology, and Physics* (Courier Dover Publications, 1976).
31. Ortega, A. & García de la Torre, J. Hydrodynamic properties of rodlike and disklike particles in dilute solution. *J. Chem. Phys.* **119**, 9914–9919 (2003).
32. Krall, A. H. & Weitz, D. A. Internal dynamics and elasticity of fractal colloidal gels. *Phys. Rev. Lett.* **80**, 778–781 (1998).
33. Friedrich, H., Frederik, P. M., de With, G. & Sommerdijk, N. A. J. M. Imaging of self-assembled structures: interpretation of TEM and cryo-TEM images. *Angew. Chem. Int. Edn* **49**, 7850–7858 (2010).
34. Hohn, M. et al. SPARX, a new environment for cryo-EM image processing. *J. Struct. Biol.* **157**, 47–55 (2007).
35. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
36. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D* **59**, 1131–1137 (2003).
37. Vuolanto, A., Uotila, S., Leisola, M. & Visuri, K. Solubility and crystallization of xylose isomerase from *Streptomyces rubiginosus*. *J. Cryst. Growth* **257**, 403–411 (2003).
38. Sleutel, M., Willaert, R., Wyns, L. & Maes, D. Kinetics and thermodynamics of glucose isomerase crystallization. *Cryst. Growth Des.* **9**, 497–504 (2009).



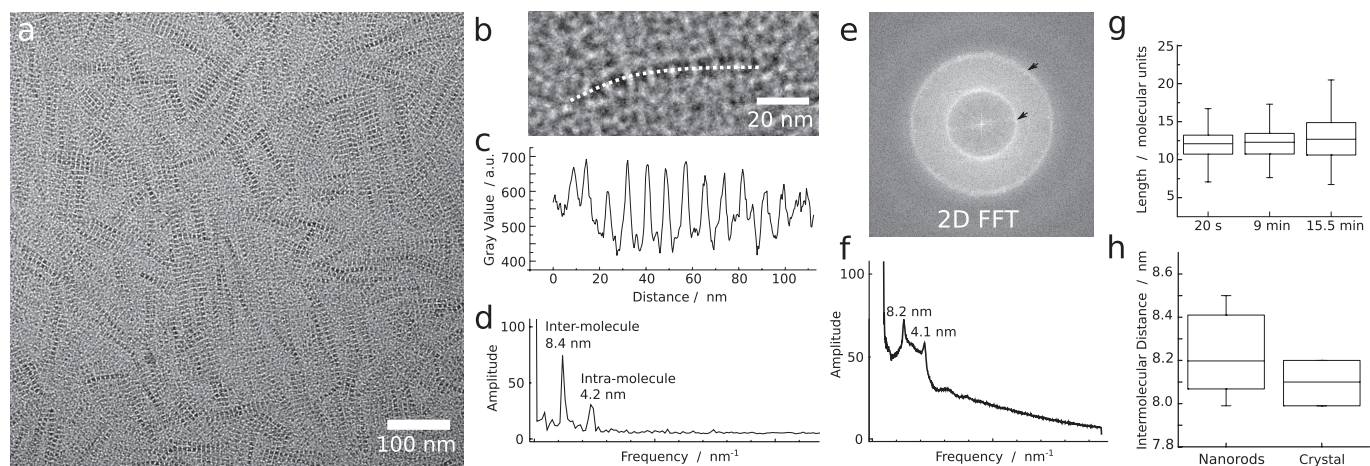
Extended Data Figure 1 | Phase diagrams for glucose isomerase.

a, b, Wide-field microscopic images of the I222 (**a**) and P2₁2₁2 (**b**) glucose isomerase (GI) polymorphs obtained with 0.8 M and 1.5 M ammonium sulfate (AS). **c**, Glucose isomerase phase diagram in ammonium sulfate ((NH₄)₂SO₄) at 22 °C (single points represent triplicate measurements), showing the solubility line C_{e,avg} (dashed line). Smaller diamonds and crosses denote smaller numbers of crystals than larger symbols. C_{e,avg} is a

mathematical average that we calculated by using the solubilities at 19 °C and 25 °C from ref. 9. **d**, Glucose isomerase phase diagram in PEG₁₀₀₀ at 22 °C, with the C_{e,avg} solubility line taken from ref. 38. The dotted lines, following the same colour code as the single points, indicate the phase boundaries in PEG₁₅₀₀. The photographs to the right are representative microscopy images at the indicated precipitant concentrations.

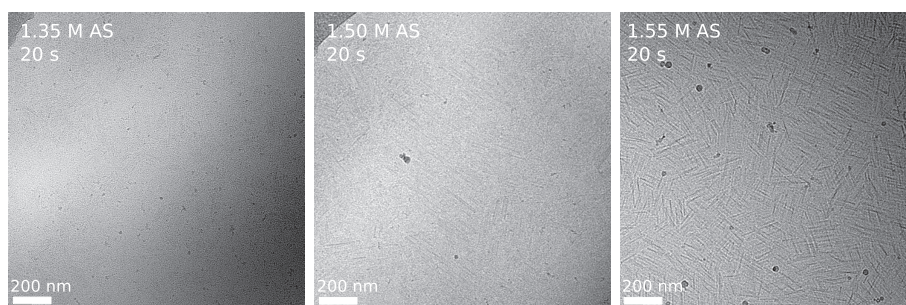


Extended Data Figure 2 | Induction time measurements. Induction time, t_{ind} , as a function of ammonium sulfate concentration. Values next to data points correspond to calculated supersaturation ($\ln C/C_e$) values, according to ref. 37.

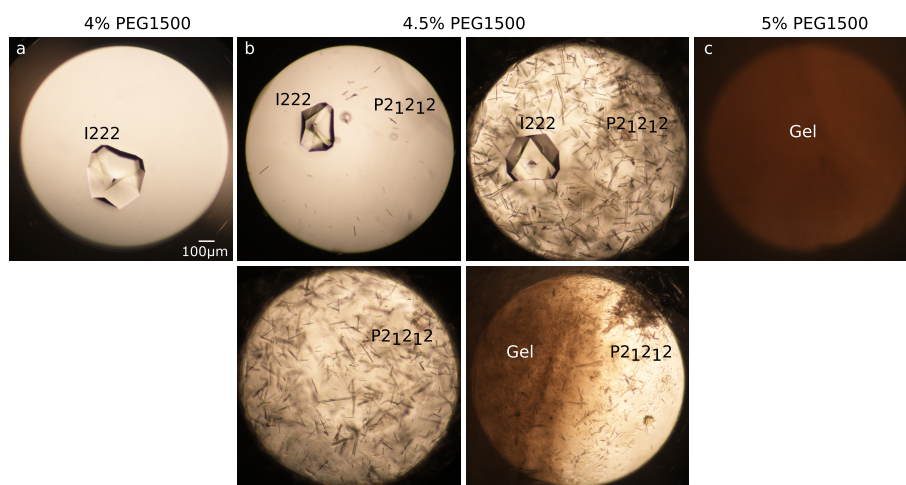


Extended Data Figure 3 | Determination of the intermolecular distance along the nanorod axis. **a**, Complete image acquired at $\times 24,000$ magnification. **b**, CryoTEM image of single nanorods. **c**, Greyscale values along the length of the dotted line in panel **a**. **d**, 1D-FFT of panel **c**, calculated using OriginPro 8.6.0. **e**, 2D-FFT image calculated using ImageJ 1.50i. **f**, Radial average of panel **e**. **g**, Nanorod length expressed

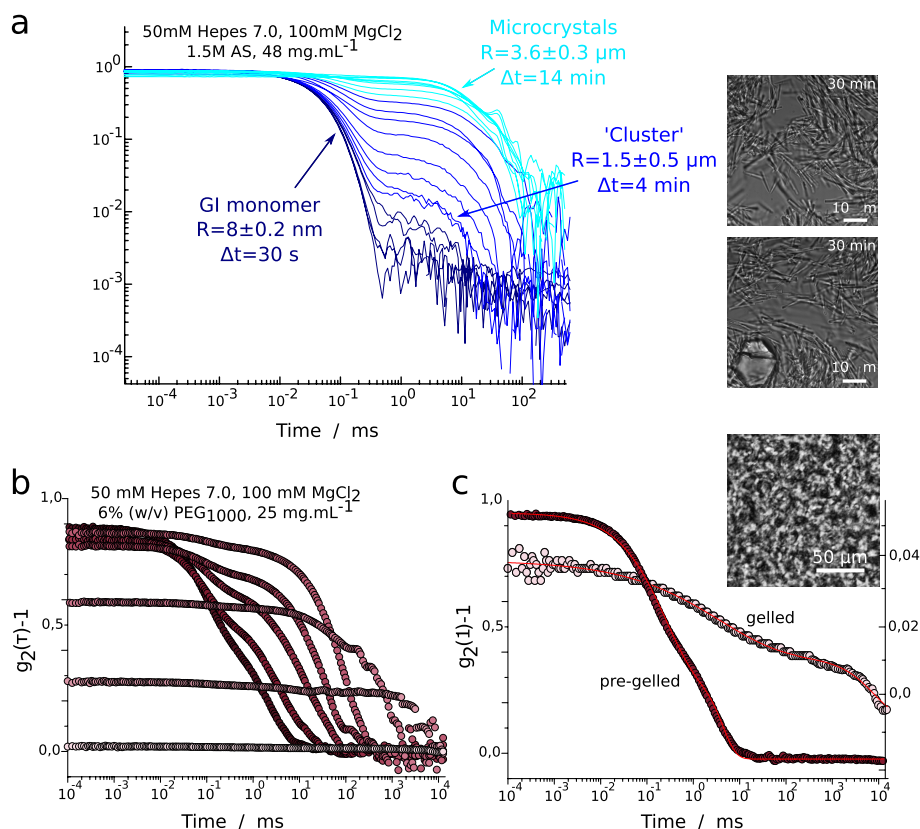
in molecular units as a function of time in 1.5 M ammonium sulfate. **h**, Intermolecular distance along the nanorod axis compared with the crystallographic distance along the c -axis of the prismatic crystals. The box range shows the 25th to 75th percentiles; the horizontal line is the median; error bars highlight the 10th and 90th percentiles.



Extended Data Figure 4 | Nanorod formation at early time points. CryoTEM images of crystallizing glucose isomerase solutions 20 seconds after protein–precipitant mixing with 1.35 M, 1.50 M or 1.55 M ammonium sulfate.

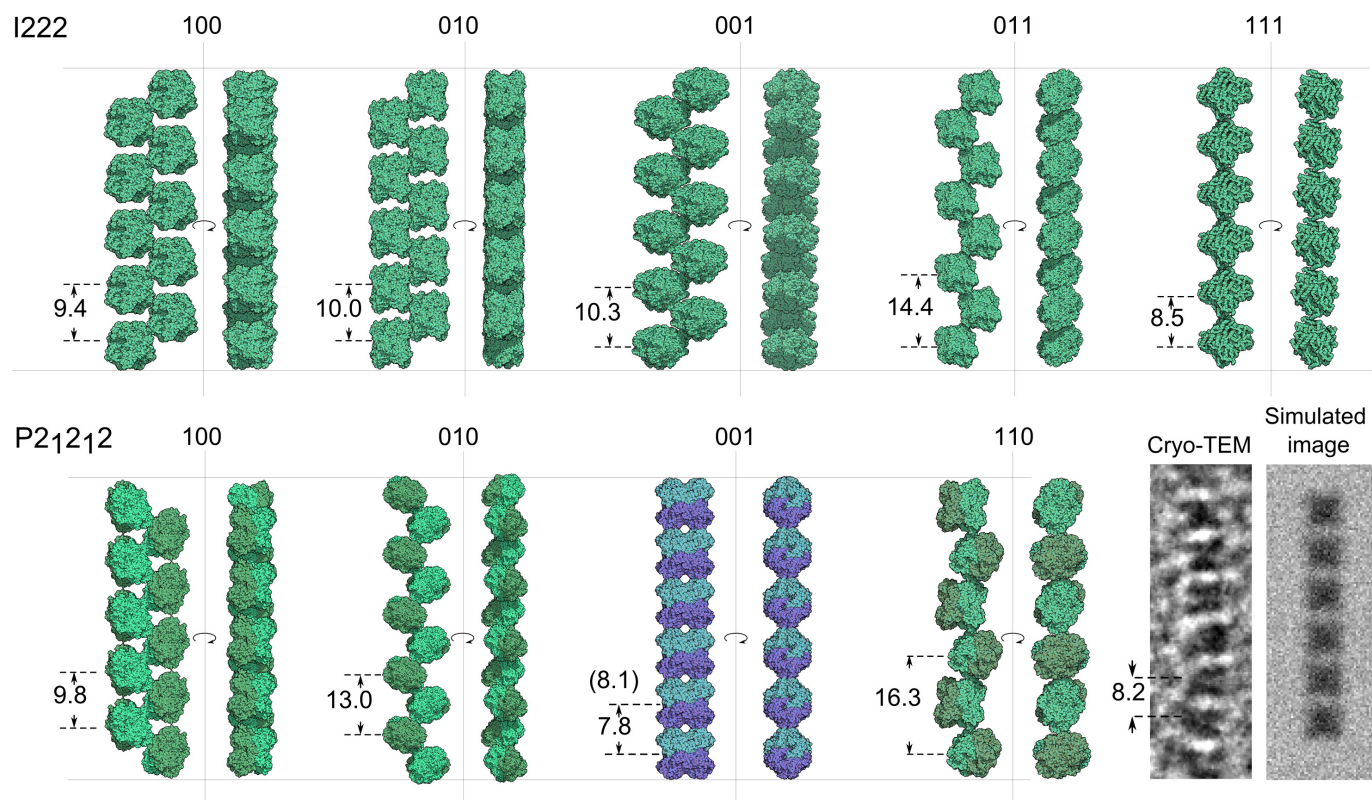


Extended Data Figure 5 | *I222/P21212/gel* coexistence point. a–c, Crystallization experiments using the microbatch-under-oil set-up, with 86 mg ml^{-1} glucose isomerase, 50 mM HEPES pH 7.0 and 100 mM MgCl_2 , and 4% (**a**), 4.5% (**b**) or 5% (**c**) (w/v) PEG_{1500} . **a**, *I222* crystals; **b**, *I222* + *P21212*, *P21212*, *P21212* and gel; **c**, gel.



Extended Data Figure 6 | Time-resolved DLS of crystallizing glucose isomerase solutions. **a**, DLS time series of a crystallizing 48 $\text{mg}\cdot\text{mL}^{-1}$ glucose isomerase solution with 50 mM HEPES pH 7.0, 100 mM MgCl_2 , 1.5 M ammonium sulfate, collected at an angle of 90° , ranging from 30 seconds to 14 minutes after protein/precipitant mixing. R , particle radius. Microscopy snapshots at the right were taken *ex situ* after

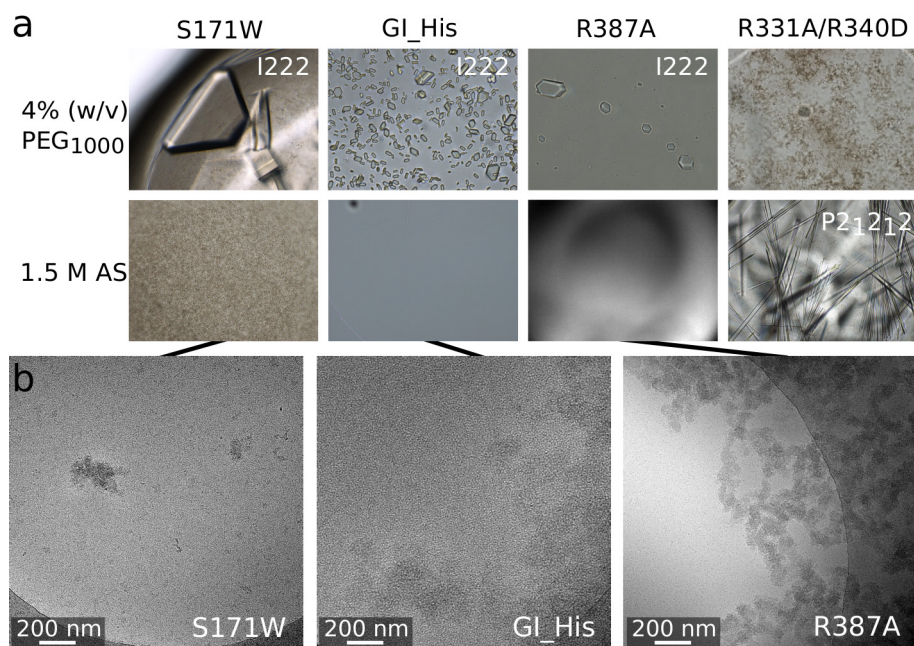
30 minutes. **b**, Time evolution (from dark to light) of the intensity correlation function of a 50 mM HEPES pH 7.0, 100 mM MgCl_2 , 6% PEG₁₀₀₀ (w/v) solution collected at an angle of 90° . **c**, Fitting of a pre-gelled (20 seconds; left-hand y -axis) and gelled (30 minutes; right-hand y -axis) sample using equations (1) and (2) respectively. Inset, wide-field microscopy image of the gelled state.



Extended Data Figure 7 | Crystallographic modelling of the nanorods.

Models of glucose isomerase nanorods in various directions, based on the unit-cell dimensions of the PDB entries 9XIA and 1OAD, and the crystallographic symmetry elements of space groups I222 and P2₁2₁2. The numbers designating intermolecular distances are in nanometres.

The number in brackets for P2₁2₁2 (001) is the value that we obtained experimentally. For reference, we compare a magnified cryoTEM image of a single nanorod and a simulated TEM projection based on the P2₁2₁2 (001) nanorod model.



Extended Data Figure 8 | Crystallization screening of glucose isomerase mutants with perturbed lattice contacts. a, Initial crystallization screening of mutants in 50 mM HEPES pH 7.0, 100 mM MgCl₂, 15 mg ml⁻¹ of glucose isomerase mutant and 4% (w/v) PEG₁₀₀₀ or 1.5 M ammonium sulfate. The mutants are S171W (with perturbed C₁

interactions), GI_His (perturbed C₁), R387A (perturbed C₂) and R331A/R340D (perturbed C₃). **b,** Cryo-TEM images of various mutants in 50 mM HEPES pH 7.0, 100 mM MgCl₂, 15 mg ml⁻¹ mutant protein and 1.5 M ammonium sulfate, 2 minutes after protein/precipitant mixing.

Extended Data Table 1 | Comparison of experimental and theoretical distances for both space groups

I222	Face	$d_{02\bar{2}}$	d_{200}	P2 ₁ 2 ₁ 2	Face	d_{001}	$d_{\bar{2}20}$
Theoretical	(011)	6,975	4,961	Theoretical	(110)	7.833	8.137
	(101)	7,151	4,703		(1 $\bar{1}$ 0)	7.833	8.137
	(110)	6,836	5,151				
Experimental	AS	6.59 ±0.15	4.72 ±0.10	Experimental	AS	8.07 ±0.12	8.36 ±0.10
	PEG	5.02	7.08				

Distances (d) are in nanometres. AS, ammonium sulfate.

Extended Data Table 2 | Lattice-contact analysis of both space groups

Space Group	Contact	H-bonds	Salt-Bridges	Patch	Δ ASA	Charges	Residues
P2 ₁ 2 ₁ 2	C1	6	0	P1	422	(1+,4-)	E167 , T170 , S171, Q172 , G173, E207, R208, E210, A344, D345, G346, L347, Q348 , A349
P2 ₁ 2 ₁ 2	C2	7	2	P2a	320	(3+,3-)	V37, E38, R41 , E70, K73, R74 , Q77 , D81
				P2b	320	(2+,2-)	A319, A322, D323, P324, E325 , R374, G385 , A386, R387 , G388
I222	C3	3	0	I1a	177	(2+,2-)	R76, D80 , <u>D81</u> , T82, G83, K85
				I1b	195	(2+,1-)	R331 , D336, R340 , P341

The residues listed in the last column have non-zero Δ ASA (\AA^2) values and are therefore considered to be (partially) buried. Residues that are involved in a hydrogen (H) bond or salt bridge with a nearest-neighbour residue are shown in bold and red, respectively. The underlined residue D81 is common to patches P2a and Ia.

Climatic control of Mississippi River flood hazard amplified by river engineering

Samuel E. Munoz^{1,2,3}, Liviu Giosan¹, Matthew D. Therrell⁴, Jonathan W. F. Remo⁵, Zhixiong Shen^{6,7}, Richard M. Sullivan^{1,8}, Charlotte Wiman¹, Michelle O'Donnell¹ & Jeffrey P. Donnelly¹

Over the past century, many of the world's major rivers have been modified for the purposes of flood mitigation, power generation and commercial navigation¹. Engineering modifications to the Mississippi River system have altered the river's sediment levels and channel morphology², but the influence of these modifications on flood hazard is debated^{3–5}. Detecting and attributing changes in river discharge is challenging because instrumental streamflow records are often too short to evaluate the range of natural hydrological variability before the establishment of flood mitigation infrastructure. Here we show that multi-decadal trends of flood hazard on the lower Mississippi River are strongly modulated by dynamical modes of climate variability, particularly the El Niño–Southern Oscillation and the Atlantic Multidecadal Oscillation, but that the artificial channelization (confinement to a straightened channel) has greatly amplified flood magnitudes over the past century. Our results, based on a multi-proxy reconstruction of flood frequency and magnitude spanning the past 500 years, reveal that the magnitude of the 100-year flood (a flood with a 1 per cent chance of being exceeded in any year) has increased by 20 per cent over those five centuries, with about 75 per cent of this increase attributed to river engineering. We conclude that the interaction of human alterations to the Mississippi River system with dynamical modes of climate variability has elevated the current flood hazard to levels that are unprecedented within the past five centuries.

Flooding of the lower Mississippi River in the spring of 2011 was among the largest discharge events since systematic measurements began in the late nineteenth century, and it caused US\$3.2 billion in agricultural losses and damages to infrastructure⁶. This and other recent flood events on the Mississippi River—including those in 2016 and 2017—have repeatedly, although controversially, been attributed to an aggressive campaign of river engineering designed and implemented over the past 150 years^{3–5}. Federally mandated efforts to reduce the impacts of flooding began in the late nineteenth century and initially relied almost exclusively on the use of artificial levees, but this strategy was revised in the wake of a particularly devastating flood in the spring of 1927 that overwhelmed the levee system⁷. The current flood management system—the Mississippi River & Tributaries Project (MR&T)—includes a series of spillways that can be opened to relieve pressure on an enlarged levee system, as well as an artificially shortened and straightened main channel that is held in place by concrete retaining walls (revetments) and isolated from most of its natural floodplain^{2,6,7}. Although these modifications are credited with protecting communities and croplands within the floodplain from inundation, artificial channelization has altered the relationship between discharge and river stage^{3,4} and accelerated the rate of land loss in the Mississippi River delta⁸, necessitating additional investments in flood mitigation infrastructure and coastal restoration⁹.

Although fluvial processes are sensitive to flood mitigation infrastructure, climate variability can also shape the dynamics of continental drainage networks, particularly over decadal to centennial timescales that are difficult to detect using short observational records^{10,11}. Precipitation and soil water storage over the Mississippi River basin are influenced by climate variability driven by sea-surface-temperature anomalies in both the Pacific and Atlantic Oceans^{12,13}. Yet establishing the natural controls on discharge extremes of the lower Mississippi has proved challenging because gauging-station measurements record a limited range of variability, particularly before major investments were made in river engineering. As a result, analyses of historical streamflow records disagree over the role that dynamical modes of climate variability play in modulating the discharge^{12,14,15}. To plan flood mitigation and other infrastructure projects, it is critical to understand the climate controls on the discharge of the lower Mississippi River, but the short length of the instrumental record limits our ability to evaluate the range of natural hydrological variability from observational data alone.

Recent advances in palaeoflood hydrology could extend the instrumental record back in time to diagnose the controls on the discharge of large alluvial rivers such as the lower Mississippi. Traditional approaches in palaeoflood hydrology, which include the use of slack-water deposits as flood event indices¹⁶, are of limited use on the low-relief landscapes that characterize the Mississippi River alluvial plain. One new approach uses the sedimentary archives held in floodplain lakes, which act as sediment traps during overbank floods, to develop continuous, quantitative and event-scale records of past flood frequency and magnitude^{17,18}. Parallel work in dendrochronology demonstrates that when trees are inundated by floodwaters they exhibit anatomical anomalies in that year's growth ring such that they provide a precise chronology of flood events that occurred during the growing season¹⁹. Together, these methodological advances provide an opportunity to evaluate interannual to multi-decadal scale trends in flood frequency and magnitude on a large alluvial river such as the lower Mississippi, before and during the era of river engineering.

Here we analyse records of individual overbank flood events derived from sedimentary and tree-ring archives from the lower Mississippi River's floodplain (Fig. 1). We collected sediment cores from the infilling thalwegs of three oxbow lakes, Lake Mary (MRY), False River Lake (FLR) and Lake Saint John (STJ), that formed by neck cut-offs of the lower Mississippi River in AD 1776, AD 1722 and roughly AD 1500, respectively²⁰ (Extended Data Figs 1–3). In these sedimentary archives, we identified individual flood events by using grain-size analysis, bulk geochemistry (from X-ray fluorescence scanning, XRF) and radiography; developed age–depth models constrained by multiple independent chronological controls (Extended Data Figs 4–6); and estimated flood magnitudes from a linear model that relates the coarse

¹Department of Geology & Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA. ²Marine Science Center, Department of Marine & Environmental Sciences, Northeastern University, Nahant, Massachusetts 01908, USA. ³Department of Civil & Environmental Engineering, Northeastern University, Boston, Massachusetts 02115, USA. ⁴Department of Geography, University of Alabama, Tuscaloosa, Alabama 35401, USA. ⁵Department of Geography and Environmental Resources, Southern Illinois University, Carbondale, Illinois 62901, USA.

⁶Department of Marine Sciences, Coastal Carolina University, Conway, South Carolina 29526, USA. ⁷Department of Geography and Planning, University of Liverpool, Liverpool L69 7ZT, UK.

⁸Department of Oceanography, Texas A&M University, College Station, Texas 77840, USA.

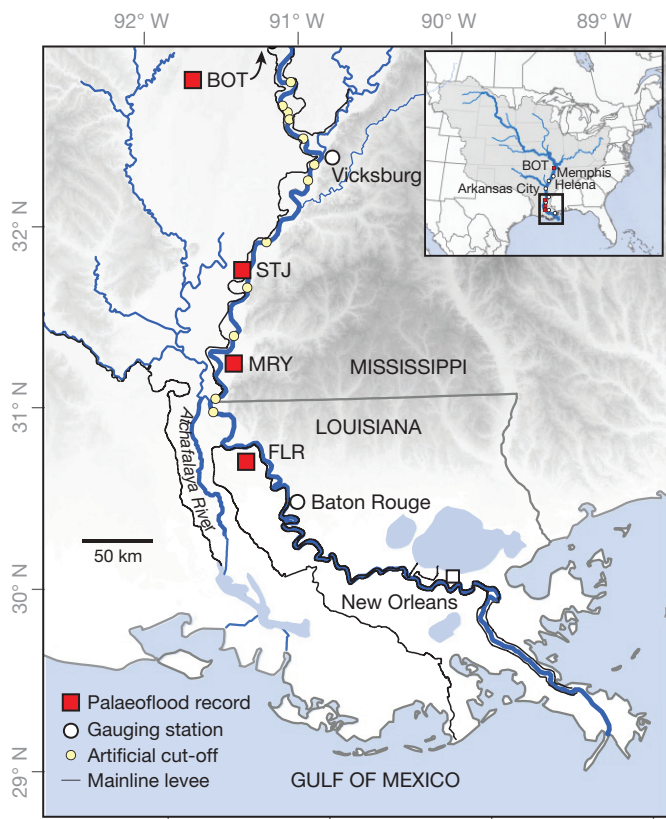


Figure 1 | The lower Mississippi River and the Mississippi River basin in North America. River engineering modifications (artificial cut-offs and levees) that contribute to channelization, the locations of palaeoflood records (FLR; MRY; STJ) and Big Oak Tree, BOT) and river gauging stations on the lower Mississippi used in this study (Memphis, Helena, Arkansas City, Vicksburg and Baton Rouge) are shown. Shaded relief shows relative topographic highs (dark shades) and lows (light shades) using the National Elevation Dataset²⁵.

grain-size component to the discharge of historical flood events¹⁸ (Extended Data Fig. 7; see Methods for details). We also include tree-ring records from the floodplain of the lower Mississippi, collected and described by ref. 21; each tree-ring series was examined for anatomical evidence of flood injury to produce a record of overbank flood events that extends back to the late seventeenth century²¹. A composite time series for flood frequency describing the number of flood events in a moving 31-year window derived from sedimentary and tree-ring archives (Fig. 2b) is highly correlated with instrumental flood frequency ($r=0.90$, $t=19.12$, effective degrees of freedom $\nu_{\text{eff}}=3.77$, $p<0.001$) for the interval of overlap, while reconstructed flood magnitudes (Fig. 2c) track trends observed in gauging-station measurements (see Supplementary Information for additional validation), indicating that the palaeoflood archives provide robust reconstructions of hydrological extremes on the lower Mississippi River beyond the period of instrumental record.

Our multi-proxy palaeoflood dataset extends the record of extremes in the discharge of the lower Mississippi River back to the early sixteenth century and demonstrates that both the frequency and magnitude of flooding have increased over the past 150 years as land use and river engineering efforts have intensified (Fig. 2). Flood frequencies and magnitudes exhibit multi-decadal oscillations that increase in amplitude around the beginning of the twentieth century such that the highest rates of overbank flooding and the largest discharge events of the past 500 years have occurred within the past century. The amplification of flood magnitudes that has occurred over the past 150 years corresponds in time with the intensification of anthropogenic modifications to the lower Mississippi River and its basin, particularly

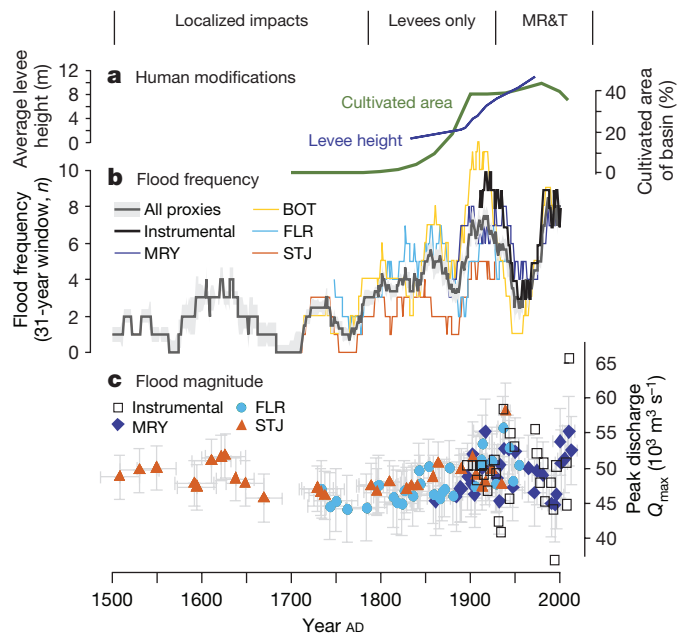


Figure 2 | Instrumental and reconstructed flood frequencies and magnitudes of the lower Mississippi River. a, Human impacts to the lower Mississippi River (MR&T refers to a major river engineering initiative): timing and intensity of agricultural land use²⁶ and river engineering. b, Flood frequencies (number of flood events in a 31-year moving window) derived from palaeoflood records, including mean and bootstrapped 2σ confidence intervals of all palaeoflood archives, and the instrumental frequency of all floods attaining major flood stage (>1.5 m above flood stage) at the Mississippi River gauging station at Baton Rouge (station number 07374000). c, Flood magnitudes derived from the sedimentary palaeoflood records, with 1σ uncertainties, and instrumental flood magnitudes for the Mississippi River gauging station at Vicksburg (station number 07289000).

the artificial channelization of the river with levees, revetments and cut-offs in the late nineteenth and early twentieth centuries^{2,7}. Yet the continued presence of multi-decadal oscillations in flood frequency and magnitude throughout the entire period of record indicates that anthropogenic modifications to the Mississippi River system are acting in concert with other factors to alter flood hazard through time.

To evaluate the role of climate variability on flood hazard, we examined the relationships between flood frequency, the El Niño–Southern Oscillation (ENSO) and the Atlantic Multidecadal Oscillation (AMO), to find that sea-surface temperature anomalies in both the Pacific and Atlantic Oceans exert a strong influence on the occurrence of lower Mississippi River floods (Fig. 3). Over the past five centuries, correlations between composite flood frequency and the frequency of El Niño events ($r=0.73$) and the AMO index ($r=-0.39$) derived from instrumental and palaeoclimate data sets are significant ($p<0.001$; see Methods for details). The strength and direction of these relationships support the hypothesis that discharge extremes on the lower Mississippi River arise through the interaction of ENSO, which influences antecedent soil moisture, with the AMO, which controls the flux of moisture from the Gulf of Mexico inland^{12,15}. Extreme precipitation events over the Mississippi River basin are associated with a stronger and more westerly position of the North Atlantic Subtropical High that is characteristic of the negative phase of the AMO^{12,13}, and these heavy precipitation events are more likely to generate discharge extremes if they fall on the saturated soils that tend to be left in the wake of El Niño events¹⁵.

Despite the strong influence of climatic variability on lower Mississippi River flood occurrence, the amplification of flood magnitudes that we observe over the past 150 years is primarily the result of human modifications to the river and its basin (Fig. 4). The magnitude

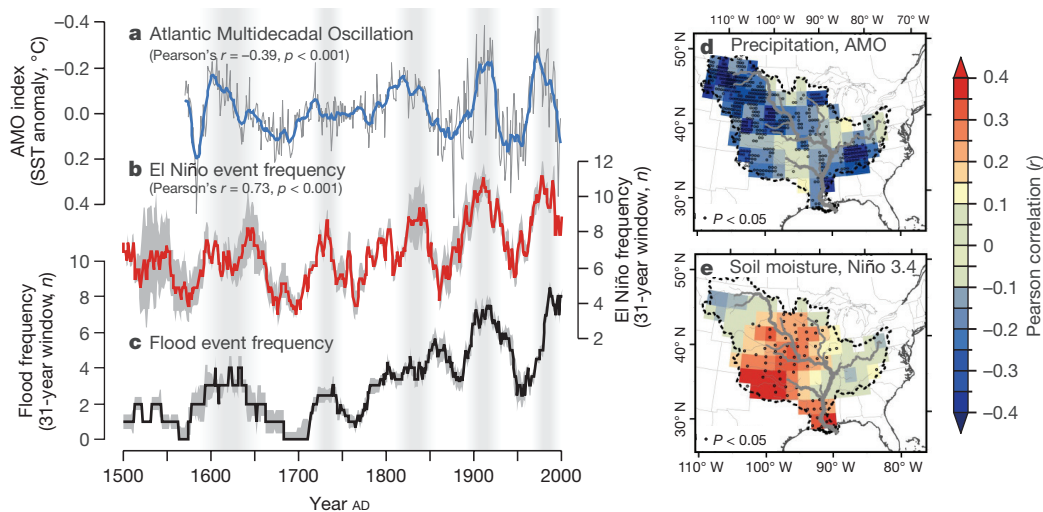


Figure 3 | Lower Mississippi River flood frequency and its relation to dominant modes of climate variability. **a**, AMO derived from instrumental²⁷ and palaeoclimate²⁸ datasets. **b**, Frequency of El Niño events (the warm phase of the ENSO) in a 31-year moving window derived from instrumental²⁷ and palaeoclimate^{28–31} data sets (mean with 2σ bootstrapped confidence interval). SST, sea surface temperature. **c**, Frequency of lower Mississippi River floods derived from palaeoflood

data (mean with bootstrapped 2σ confidence interval). **d**, Correlation field of monthly precipitation³² with the AMO²⁷ (AD 1901–2014) smoothed with a common 121-month filter. **e**, Correlation field of monthly Palmer Drought Severity Index³³ with the Niño 3.4 index²⁷ (AD 1948–2011). Correlation fields are interpolated to a common $2^\circ \times 2^\circ$ grid, and individual points with significant correlations at the $P < 0.05$ level are marked with a hollow circle.

of the 100-year flood (Q_{100} ; a flood with a 1% chance of exceedance in any year) estimated from gauging-station measurements (AD 1897–2015) is $(20 \pm 7)\%$ larger than Q_{100} for the period before major human impacts to the river and its basin (AD 1500–1800), as estimated from the palaeoflood data (see Methods for details). To identify the influence of human activities on this observed increase in Q_{100} , we use a linear model that relates peak discharge to the AMO index over the period before major human impacts to the river, AD 1500–1800 ($R^2 = 0.35$, degrees of freedom $\nu = 18$, $p < 0.01$) and use this model to predict flood magnitudes over the entire period of record. This ‘climate-only’ regression predicts that, in the absence of human modifications to the land surface, Q_{100} would have increased by only $(5 \pm 6)\%$ over the same period, accounting for only about 25% of the observed increase in Q_{100} and implying that the remainder (about 75%) of this elevated flood hazard is the result of human modifications to the river and its basin.

The timing and nature of the amplification of flood magnitudes at the onset of the twentieth century strongly imply that it reflects the transformation of a freely meandering alluvial river to an artificially confined channel, because the confinement of flood flows to a levee-defined floodway can speed up the downstream propagation of a flood wave and increase peak discharge for a given flood²². The establishment of widespread agricultural activity in the Mississippi River basin occurred in the nineteenth century, before the divergence of the observed and ‘climate-only’ flood magnitudes, indicating a secondary and possibly lagged influence of agricultural expansion²³ on flood magnitudes relative to that of river engineering. In short, this analysis identifies artificial channelization of the lower Mississippi River, and its effects on the river’s gradient, channel area and flow velocity^{2,7}, as having significantly increased the discharge of a given flood event relative to pre-engineering conditions.

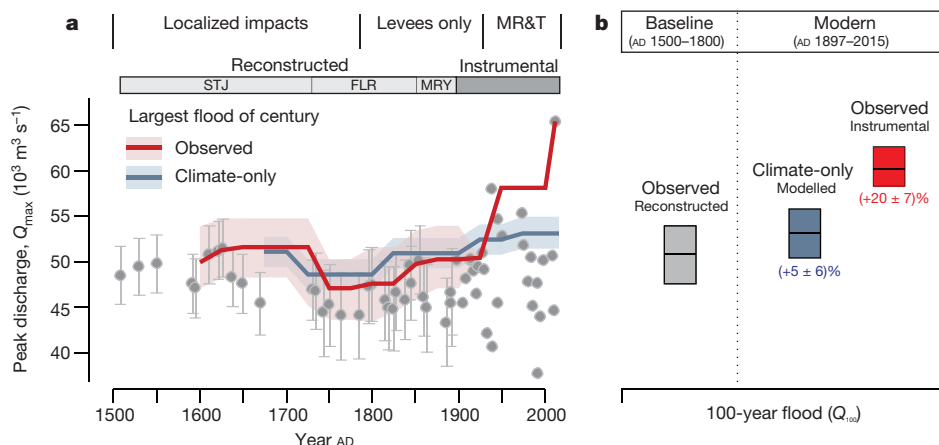


Figure 4 | Attribution of the observed increase in flood magnitudes over the past five centuries. **a**, Composite peak discharges from palaeoflood archives and the instrumental record from Vicksburg. The red line indicates observed trends in the largest flood of the century in a moving window; the blue line indicates trends under ‘climate-only’ conditions, estimated from a statistical model (see text for details). Both lines are shown with 1σ confidence intervals. Instrumental peak discharge estimates are reported without uncertainty and are therefore

plotted without confidence intervals. **b**, Comparison of the 100-year flood observed during the baseline period (AD 1500–1800, before major human modifications to the Mississippi River and its basin; grey boxplot) with that estimated using a statistical model under ‘climate-only’ conditions (blue boxplot) and observed (red boxplot) during the modern period of instrumental record (AD 1897–2015). Boxplots show mean (centre line) and 1σ confidence intervals (box top and bottom) for Q_{100} estimates.

Our main finding—that river engineering has elevated flood hazard on the lower Mississippi to levels that are unprecedented within the past five centuries—adds to a growing list of externalized costs associated with conventional flood mitigation and navigation projects, including a reduction in a river's ability to convey flood flows^{3,4}, the acceleration of coastal land loss⁸ and hypoxia²⁴. Despite the societal benefits that these major infrastructure projects convey⁶, the costs associated with maintaining current levels of flood protection and navigability will continue to grow at the expense of communities and industries situated in the river's floodplain and its delta. For those interested in improving seasonal and longer-term forecasts of flood hazard or management strategies that reconnect the river with its floodplain, the Mississippi River's discharge of freshwater—and by extension the flux of sediment, nutrients and pollutants—to its outlet should be viewed as highly sensitive both to anthropogenic modifications to the basin and to variability of the global climate system.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 June 2017; accepted 31 January 2018.

- Meybeck, M. Global analysis of river systems: from Earth system controls to Anthropocene syndromes. *Phil. Trans. R. Soc. Lond. B* **358**, 1935–1955 (2003).
- Smith, L. M. & Winkley, B. R. The response of the lower Mississippi River to river engineering. *Eng. Geol.* **45**, 433–455 (1996).
- Criss, R. E. & Shock, E. L. Flood enhancement through flood control. *Geology* **29**, 875–878 (2001).
- Pinter, N., Jemberie, A. A., Remo, J. W., Heine, R. A. & Ickes, B. S. Flood trends and river engineering on the Mississippi River system. *Geophys. Res. Lett.* **35**, L23404 (2008).
- Watson, C. C., Biedenharn, D. S. & Thorne, C. R. Analysis of the impacts of dikes on flood stages in the Middle Mississippi River. *J. Hydraul. Eng.* **139**, 1071–1078 (2013).
- Camillo, C. A. *Divine Providence: The 2011 Flood in the Mississippi River and Tributaries Project* (Mississippi River Commission, 2012).
- Remo, J. W. F. *Fishery Resources, Environment, and Conservation in the Mississippi and Yangtze (Changjiang) River Basins* Ch. 11 (American Fisheries Society, 2016).
- Blum, M. D. & Roberts, H. H. Drowning of the Mississippi Delta due to insufficient sediment supply and global sea-level rise. *Nat. Geosci.* **2**, 488–491 (2009).
- Louisiana Coastal Protection and Restoration Authority. *Louisiana's Comprehensive Master Plan for a Sustainable Coast* (Coastal Protection and Restoration Authority of Louisiana, 2017).
- Aalto, R. *et al.* Episodic sediment accumulation on Amazonian flood plains influenced by El Niño/Southern Oscillation. *Nature* **425**, 493–497 (2003).
- Darby, S. E. *et al.* Fluvial sediment supply to a mega-delta reduced by shifting tropical-cyclone activity. *Nature* **539**, 276–279 (2016).
- Enfield, D. B., Mestas-Núñez, A. M. & Trimble, P. J. The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US. *Geophys. Res. Lett.* **28**, 2077–2080 (2001).
- Hu, Q., Feng, S. & Oglesby, R. J. Variations in North American summer precipitation driven by the Atlantic Multidecadal Oscillation. *J. Clim.* **24**, 5555–5570 (2011).
- Rogers, J. C. & Coleman, J. S. Interactions between the Atlantic Multidecadal Oscillation, El Niño/La Niña, and the PNA in winter Mississippi valley stream flow. *Geophys. Res. Lett.* **30**, (2003).
- Munoz, S. E. & Dee, S. G. El Niño increases the risk of lower Mississippi River flooding. *Sci. Rep.* **7**, <https://doi.org/10.1038/s41598-017-01919-6> (2017).
- Baker, V. R. Paleoflood hydrology and extraordinary flood events. *J. Hydrol.* **96**, 79–99 (1987).
- Munoz, S. E. *et al.* Cahokia's emergence and decline coincided with shifts of flood frequency on the Mississippi River. *Proc. Natl Acad. Sci. USA* **112**, 6319–6324 (2015).
- Toonen, W. H. J., Winkels, T. G., Cohen, K. M., Prins, M. A. & Middelkoop, H. Lower Rhine historical flood magnitudes of the last 450 years reproduced from grain-size measurements of flood deposits using end member modelling. *Catena* **130**, 69–81 (2015).
- St. George, S. & Nielsen, E. Signatures of high-magnitude nineteenth-century floods in *Quercus macrocarpa* tree rings along the Red River, Manitoba, Canada. *Geology* **28**, 899–902 (2000).
- Fisk, H. N. *Geological Investigation of the Alluvial Valley of the Lower Mississippi River* (Mississippi River Commission, 1945).
- Therrell, M. D. & Bialecki, M. B. A multi-century tree-ring record of spring flooding on the Mississippi River. *J. Hydrol.* **529**, 490–498 (2015).
- Jacobson, R. B., Lindner, G. & Bitner, C. The role of floodplain restoration in mitigating flood risk, lower Missouri River, USA. *Geomorphic Approaches to Integrated Floodplain Management of Lowland Fluvial Systems in North America and Europe* 203–243 (Springer, 2015).
- Trimble, S. W. Decreased rates of alluvial sediment storage in the Coon Creek Basin, Wisconsin, 1975–93. *Science* **285**, 1244–1246 (1999).
- Rabalais, N. N. *et al.* Dynamics and distribution of natural and human-caused hypoxia. *Biogeosciences* **7**, 585–619 (2010).
- Gesch, D. *et al.* The National Elevation Dataset. *Photogramm. Eng. Remote Sensing* **68**, 5–32 (2002).
- Klein Goldewijk, K., Beusen, A., Doelman, J. & Stehfest, E. New anthropogenic land use estimates for the Holocene; HYDE 3.2. *Earth Syst. Sci. Data Discuss.* <https://doi.org/10.5194/essd-2016-58> (2016).
- Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmos.* **108**, 4407 (2003).
- Braganza, K., Gergis, J. L., Power, S. B., Risbey, J. S. & Fowler, A. M. A multiproxy index of the El Niño–Southern Oscillation, AD 1525–1982. *J. Geophys. Res. Atmos.* **114**, D05106 (2009).
- Gergis, J. L. & Fowler, A. M. A history of ENSO events since AD 1525: implications for future climate change. *Clim. Change* **92**, 343–387 (2009).
- Li, J. *et al.* Interdecadal modulation of El Niño amplitude during the past millennium. *Nat. Clim. Chang.* **1**, 114–118 (2011).
- McGregor, S., Timmermann, A. & Timm, O. A unified proxy for ENSO and PDO variability since 1650. *Clim. Past* **6**, 1–17 (2010).
- Schneider, U. *et al.* Evaluating the hydrological cycle over land using the newly-corrected precipitation climatology from the Global Precipitation Climatology Centre (GPCC). *Atmosphere* **8**, 52–69 (2017).
- Vose, R. S. *et al.* Improved historical temperature and precipitation time series for U.S. climate divisions. *J. Appl. Meteorol. Climatol.* **53**, 1232–1251 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Colman, S. G. Dee, K. Lotterhos, S. P. Muñoz, W. H. J. Toonen, G. C. Trussell and T. Webb III for discussion and comments, and M. Besser, D. Carter, J. Elsenbeck, K. Esser, A. LaBella and J. Nienhuis for field and/or laboratory assistance. Seed funding for this project was provided to L.G. and J.P.D. by the Coastal Ocean Institute of WHOI. Support for S.E.M. was provided by the Postdoctoral Scholar Program of the Woods Hole Oceanographic Institution (WHOI). Additional support to S.E.M. and L.G. was provided by the Ocean and Climate Change Institution of WHOI. Support for M.D.T. and J.W.F.R. was provided by the US National Science Foundation Geography and Spatial Science Program (award number BSC1359801). This is contribution no. 362 from the Marine Science Center at Northeastern University.

Author Contributions L.G. and J.P.D. initiated the project. S.E.M., L.G., M.D.T., J.W.F.R., Z.S. and J.P.D. conceived the ideas, designed the study and interpreted the results. M.D.T. provided dendrochronological data. J.W.F.R. provided historical discharge and geospatial data. Z.S. performed OSL dating. S.E.M., L.G., R.M.S., C.W. and M.O. collected sedimentary archives and/or performed laboratory analyses. S.E.M. wrote the manuscript with contributions from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.E.M. (s.munoz@northeastern.edu).

Reviewer Information *Nature* thanks P. Hudson, S. St George and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Instrumental streamflow data. We obtained daily stage data for Mississippi River gauges at Vicksburg (station number 07289000) and Baton Rouge (07374000) from the United States Army Corps of Engineers (USACE) and the United States Geological Survey (USGS). Discharges for the Vicksburg, Memphis (07032000), Helena (07047970), Arkansas City (07146500) and Baton Rouge gauges were compiled from multiple sources. For the early instrumental record (pre-1927), peak discharges and measured discharges were compiled from historical documents^{34,35}. In the few cases in which annual peak discharges were not recorded during this period, we used the measured discharges to create rating curves from which to determine the peak discharge for the annual peak stage. Discharge data after AD 1927 were acquired either from the USACE or from the USGS. The discharge record at Vicksburg is the longest and most continuous of the available discharge records, and its peak annual discharge is highly correlated ($r > 0.86$, $p < 0.01$) with that of other lower Mississippi River gauging stations in the study area (see Supplementary Information) and was thus used to reconstruct flood magnitudes from the sedimentary archives.

Sedimentary archives. We collected sediment cores from the infilling thalwegs of MRY, FLR and STJ with a rod-driven vibracore system in July 2012 and March 2016 (Extended Data Figs 1–3). For each core, we collected a replicate drive using a 7.5-cm-diameter polycarbonate piston corer to ensure recovery of an intact sediment/water interface. The targeted lakes were selected because the lateral position of the active channel near the lake's arm has remained relatively stable from the time of cut-off to the mid-twentieth century²⁰. We cannot eliminate the possibility that minor lateral and/or vertical channel migration has occurred near these lakes since the time of cut-off, but we reduce the influence of this potential bias on our analysis by (i) using a low-pass filter on the grain-size data (see below) and (ii) validating the resulting flood frequency and magnitude data sets against the instrumental record (see Supplementary Information). At FLR and STJ, mainline levees of the MR&T have inhibited the deposition of fluvial sediment in the lake during overbank floods after about AD 1950 and 1937, respectively; MRY is not protected by artificial levees and it continues to be inundated during overbank floods. Oxbow lakes can continue to exchange water and sediment with the main channel when the river is below flood stage³⁶ to create high rates of fine-grained 'background sedimentation' that differs in texture and composition from the coarser material that is mobilized during high-magnitude flood events. Cores were collected along an arm of the oxbow lakes at locations proximal to the 'plug' that separates the active channel from the lake to maximize the contrast between background and flood event sediments. Core locations at each site were targeted based on bathymetric surveys before core collection.

Cores were transported back to the Woods Hole Oceanographic Institution (WHOI) where they were split, described and photographed. Archived core halves were subjected to high-resolution XRF (4,000 μm resolution) and radiography (200 μm resolution) in an ITRAX core scanner housed at WHOI. For grain-size analysis, sediment sub-samples at continuous 1-cm intervals were dispersed in water using a vortex mixer before 5 s sonication and analysis in a Beckman Coulter LS 13 320 laser diffraction particle-size analyser; randomly selected replicate samples showed a $< 1\%$ volume difference in any detector. Complex, multi-modal grain-size distributions were modelled as mixtures of discrete, simple distributions and decomposed using end-member calculations into four representative populations, or end-members (EMs), that were considered geologically meaningful, using the EMMAgeo package run in RStudio. The score of each sample on the coarsest end-members (EM1), representing deposition of bedload during overbank floods¹⁸, was normalized with a low-pass (41-cm) moving minimum filter to remove long-term trends in sediment composition caused by local geomorphic processes. We then identified potential flood deposits as normalized EM1 scores that exceeded a high-pass (11-cm) moving mean with a 0.1 EM1 score threshold, and we verified identified peaks against the XRF and radiography (Extended Data Figs 4–6).

To estimate flood magnitudes from the sediment records, we used the method of ref. 18 and developed linear models that describe the normalized EM1 scores as a function of historical flood event discharge at the Mississippi River gauging station at Vicksburg. Using this, we assigned each flood deposit to a historical flood event approximating 'major flood stage' as defined by the USGS at a nearby gauging station, in stratigraphic order, and within the 2σ age estimate for the deposit (Extended Data Fig. 7). The requirement for flood deposits to be assigned to historical floods in stratigraphic order eliminated ambiguity in cases in which more than one historical flood fell within a deposit's 2σ age estimate. There were no cases for which a flood deposit could not be assigned to a historical flood within the period of instrumental observations (AD 1897–2015), but there were three cases at FLR (AD 1944, 1929 and 1920) and two cases at STJ (AD 1920 and 1913) for which a major historic flood did not leave an identifiable flood deposit. These 'missing' flood deposits are rare and occurred during periods of high flood frequency, and they may reflect reduced sediment availability³⁷ during these events.

The sedimentary record reconstructs peak annual discharge at the Vicksburg gauge, not at individual site locations.

We developed age–depth models using Bacon v.2.2³⁸, a Bayesian age–depth modelling program, informed by multiple independent dating techniques (see Supplementary Information), including: (i) ^{137}Cs and ^{210}Pb activity in desiccated and powdered bulk sediment samples in a Canberra GL2020RS well detector for low-energy germanium gamma radiation, for which we used the constant rate of supply model³⁹ to estimate the age of a sampled depth; (ii) radiocarbon (^{14}C) dating via accelerator mass spectrometry of a terrestrial plant macrofossil at the National Ocean Sciences Accelerator Mass Spectrometers facility at WHOI, calibrated using the IntCal13 curve embedded in Bacon; (iii) optically stimulated luminescence (OSL) dating with the fast component of silt-sized quartz⁴⁰ using a Risø DA-15 B/C luminescence reader at the University of Liverpool, UK; (iv) core tops as the date of collection and, when appropriate, the age of lake formation²⁰ as the core bottom. Sedimentation rate priors were increased to near-instantaneous rates through thick ($> 20\text{ cm}$) flood deposits¹⁷.

Tree-ring records. Tree-ring samples from 33 living and 2 dead oak (*Quercus lyrata* and *Q. macrocarpa*) trees were collected from Big Oak Tree State Park (BOT) in southeast Missouri²¹. One to four core samples were extracted from each tree at or below breast height (about 1.4 m) using a 5-mm-diameter Swedish increment borer. Cross-sections from dead trees were collected as close to the base of the tree as possible. All samples were absolutely cross-dated using the skeleton-plot method of dendrochronology. Tree-ring widths were measured on a stage micrometer to a nominal resolution of 0.001 mm. We crosschecked the accuracy of our visual dating using the computer program COFECHA. We visually determined flood-ring years by examining each tree-ring series for any evidence of flood injury consistent with the anomalous anatomical features caused by flooding as described by previous flood-ring studies¹⁹. Additional characteristics used in our identification included 'jumbled ranks' or 'additional ranks' of early wood vessels or zones of 'extended earlywood' and disorganized flame parenchyma as well as 'offset' early wood ranks¹⁹. We used the same criteria as ref. 21 to identify flood events (that is, a year in which more than 10% of sampled trees exhibited signs of flood injury) as this threshold encompasses all historic floods that attained major flood stage and occurred during the growing season²¹.

Historical climate and palaeoclimate data. Historical (late nineteenth century to present) indices of ENSO and AMO²⁷ were extended back to the sixteenth century with annual palaeoclimate reconstructions of ENSO^{28–31} and AMO⁴¹. To compare the ENSO series, we identified El Niño events in the historical Niño 3.4 index as periods of five consecutive overlapping 3-month windows at or above $+0.5^\circ\text{C}$, and as years with anomalies of more than $+0.5^\circ\text{C}$ in the palaeoclimate series. We then derived El Niño event frequencies using a 31-year moving window on each record, and we computed the mean of the historical and all palaeoclimate El Niño frequencies and bootstrapped 2σ confidence intervals using the *boot* function in RStudio. For the composite AMO series, we used the detrended historical AMO index²⁷ back to AD 1871, and then transitioned to a palaeoclimate AMO reconstruction⁴¹ to AD 1572. We sampled this composite AMO index at the median age probability of the 20 palaeofloods that occurred between AD 1500–1800, and used these data to develop a linear model (using the *lm* function in RStudio) that relates peak discharge from the AMO index; the El Niño frequency timeseries was not a significant predictor of flood magnitudes, presumably because Pacific sea-surface temperatures do not control the inland flux of Gulf of Mexico moisture that triggers high-magnitude discharge events¹⁵, so only the AMO index was used to statistically estimate flood magnitudes under 'climate-only' conditions. The AMO is detrended to remove recent warming of North Atlantic sea surface temperatures, so the 'climate-only' estimates of Q_{100} do not consider the potential effects of recent greenhouse warming on flood magnitudes—although we note that the inverse relationship between AMO and Mississippi River flood magnitudes implies that warming of North Atlantic sea-surface temperatures would act to suppress flood magnitudes. When evaluating the significance of Pearson correlations between climate and hydrological time-series that exhibited high degrees of serial autocorrelation, we estimated the effective degrees of freedom with the following relation⁴²:

$$\nu_{\text{eff}} = N(1 - \varphi_x \varphi_y) / (1 + \varphi_x \varphi_y) \quad (1)$$

where N is the number of independent samples, and φ_x and φ_y are the lag-1 autocorrelation coefficients of time series x and y respectively.

Flood hazard attribution. The magnitude of Q_{100} was estimated both empirically and through statistical modelling. The sedimentary palaeoflood archives record major flood events over periods greater than 100 years, and are suitable for estimating recurrence intervals empirically through the relation:

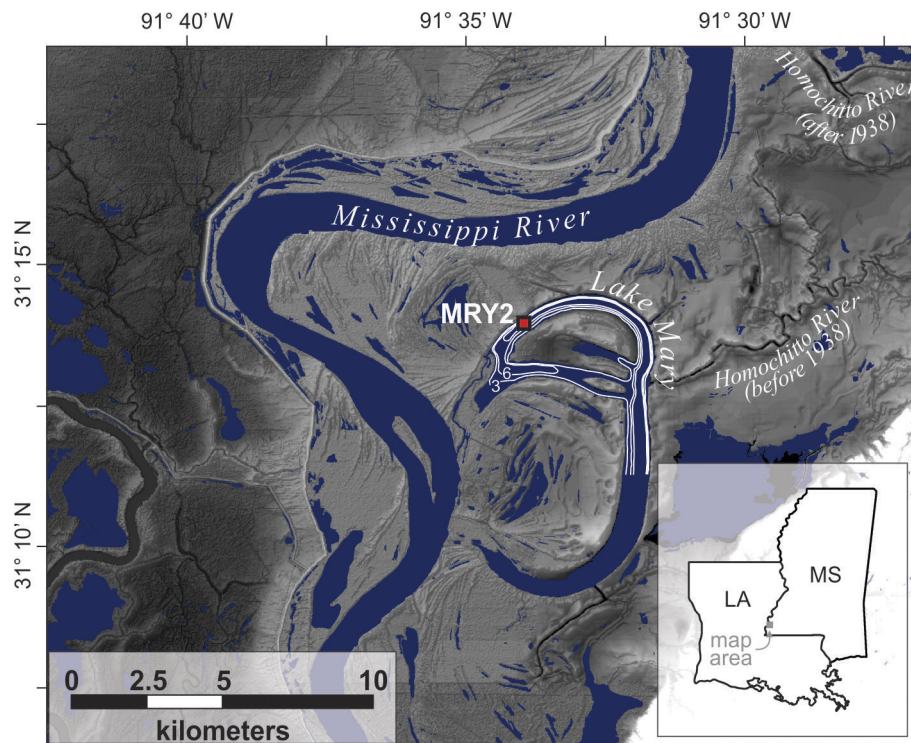
$$t_r = (n + 1) / m \quad (2)$$

where t_r is the recurrence interval (the inverse of t_r is the probability that the event magnitude will be exceeded in any one year), n is the number of years in the window being considered, and m is the number of recorded occurrences of the event being considered. The same approach was used to estimate Q_{100} in the statistically modelled 'climate-only' peak annual discharges derived from palaeoclimate and historical climate records. The instrumental record at the Vicksburg gauge provides a measurement for peak annual discharge in every year, but is relatively short, so the modern Q_{100} was estimated statistically by fitting a log Pearson type III distribution to the data set following standard protocols outlined by the United States Interagency Advisory Committee of Water Data⁴³ for instrumental hydrological data sets. We compared the observed Q_{100} baseline (AD 1500–1800) with the observed and 'climate-only' Q_{100} estimates for the modern period (AD 1897–2015) and attributed the proportion of the observed change that was not explained by the 'climate-only' estimates to human alterations to the river channel and basin. The modern Q_{100} estimated empirically from sedimentary records and the modern Q_{100} estimated by fitting a generalized extreme value distribution to the instrumental data both fall within the 1σ confidence intervals of the modern Q_{100} estimated by fitting a log Pearson type III to the instrumental record (see Supplementary Information), indicating that our findings are robust to different estimations of flood hazard.

Data availability statement. The datasets generated by this study are available as Supplementary Data.

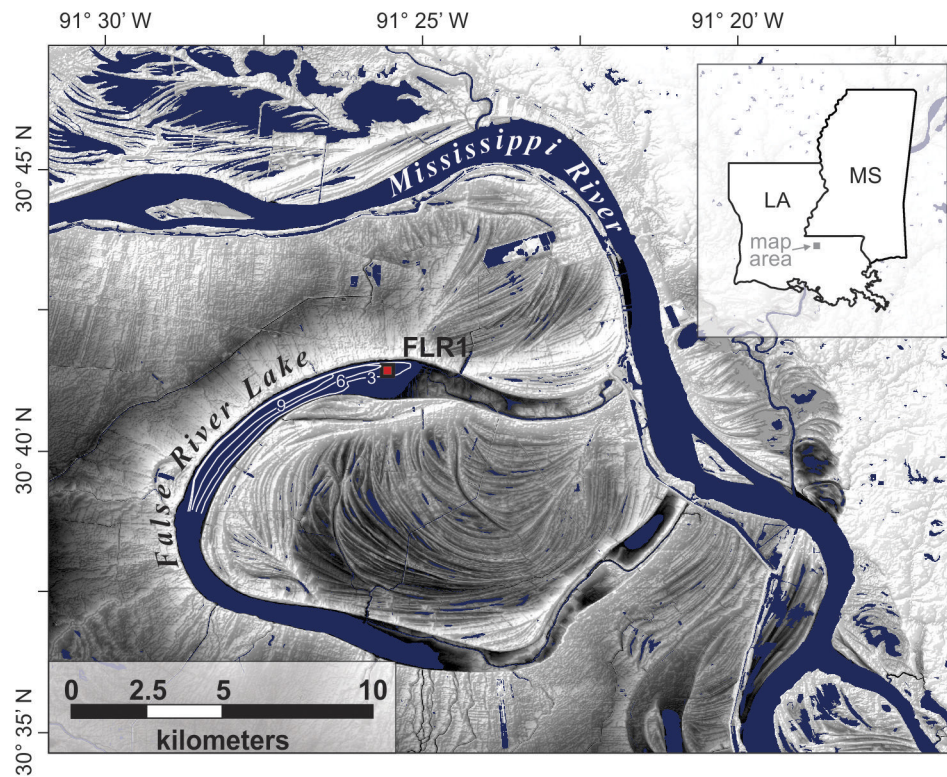
Code availability. The R code used to produce the figures in this paper is available from the corresponding author on reasonable request.

34. Mississippi River Commission. *Results of the Discharge Observations Mississippi River and its Tributaries and Outlets, 1838–1923* (Mississippi River Commission, 1925).
35. Mississippi River Commission. *Results of the Discharge Observations Mississippi River and its Tributaries and Outlets, 1924–1930* (Mississippi River Commission, 1931).
36. Hudson, P. F., Sounny-Slitine, M. A. & LaFevor, M. A new longitudinal approach to assess hydrologic connectivity: embanked floodplain inundation along the lower Mississippi River. *Hydrol. Processes* **27**, 2187–2196 (2013).
37. Heitmüller, F. T., Hudson, P. F. & Kesel, R. H. Overbank sedimentation from historic ad 2100 flood along the lower Mississippi River, USA. *Geology* **45**, 107–110 (2017).
38. Blaauw, M. & Christen, J. A. Flexible paleoclimate age–depth models using an autoregressive gamma process. *Bayesian Anal.* **6**, 457–474 (2011).
39. Appleby, P. G. & Oldfield, F. The calculation of lead-210 dates assuming a constant rate of supply of unsupported ^{210}Pb to the sediment. *Catena* **5**, 1–8 (1978).
40. Shen, Z. & Lang, A. Quartz fast component optically stimulated luminescence: towards routine extraction for dating applications. *Radiat. Meas.* **89**, 27–34 (2016).
41. Gray, S. T., Graumlich, L. J., Betancourt, J. L. & Pederson, G. T. A tree-ring based reconstruction of the Atlantic Multidecadal Oscillation since 1567 AD. *Geophys. Res. Lett.* **31**, L12205 (2004).
42. Dawdy, D. & Matias, N. *Statistical and Probability Analysis of Hydrologic Data, Part III: Analysis of Variance, Covariance and Time Series* (McGraw-Hill, 1964).
43. Interagency Advisory Committee on Water Data. *Guidelines for Determining Flood-Flow Frequency: Bulletin 17B of the Hydrology Subcommittee* (United States Geological Survey, 1982).



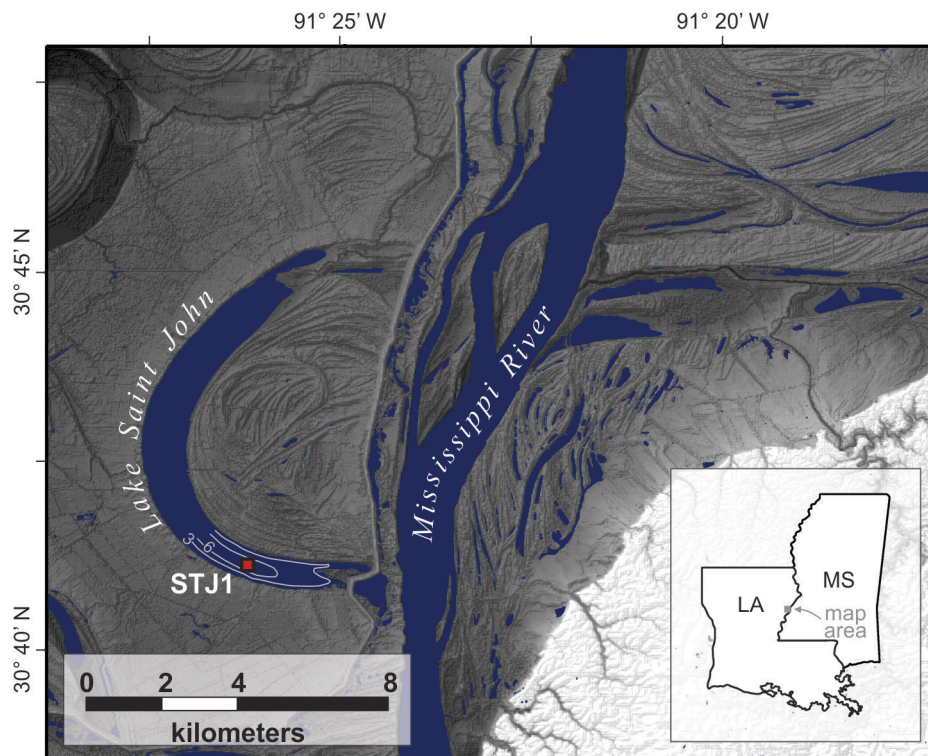
Extended Data Figure 1 | Location of Lake Mary, Mississippi (MRY) and sediment core (MRY2) used in this study. Lake Mary is an oxbow lake that formed via neck cut-off of the lower Mississippi River in AD 1776²⁰ and is situated inside the modern floodway such that it continues to

be inundated during overbank floods. Bathymetric contours (white) given in metres. Shaded relief shows relative topographic lows (dark shades) and highs (light shades) according to the National Elevation Dataset²⁵.



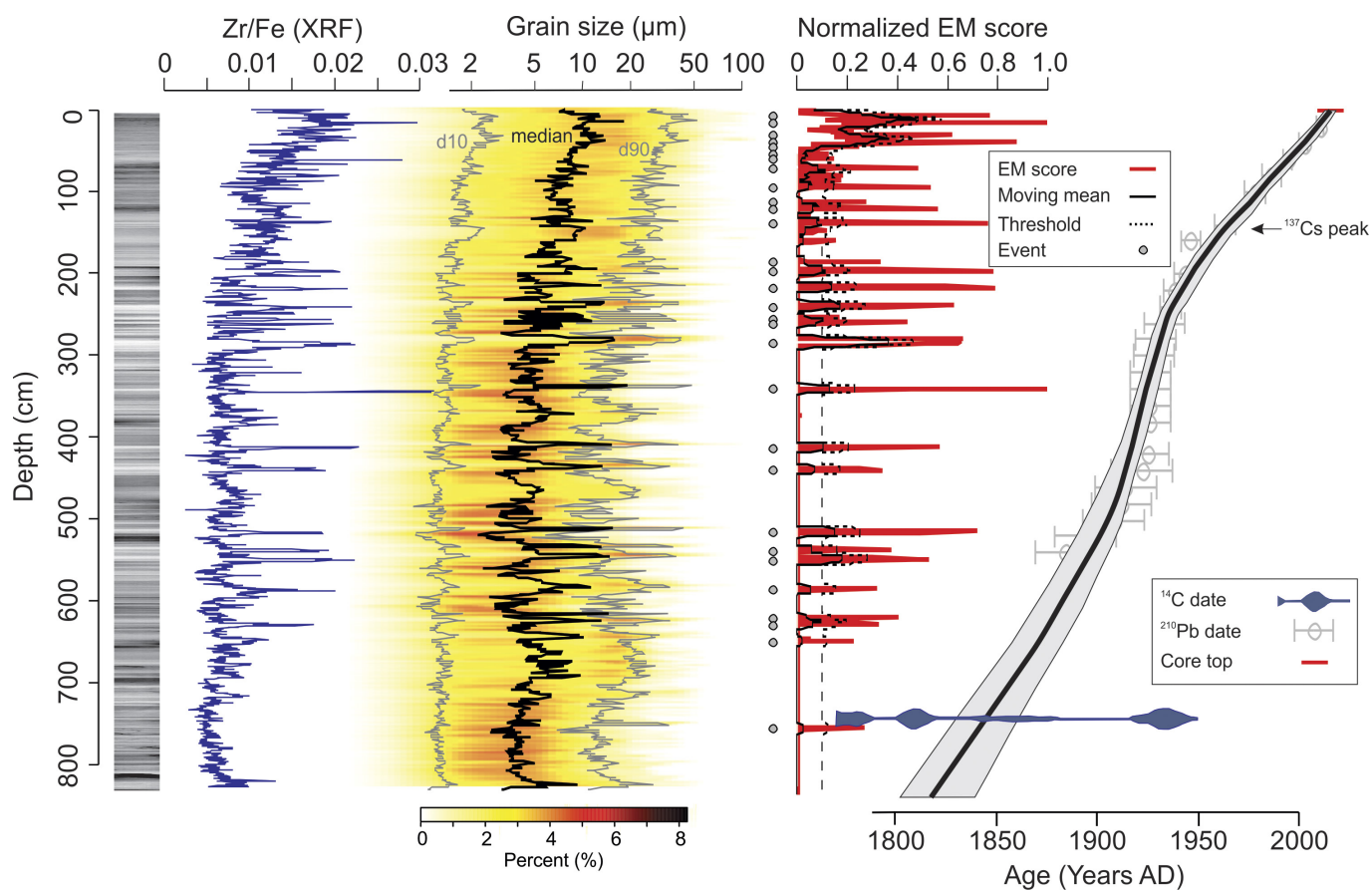
Extended Data Figure 2 | Location of False River Lake, Louisiana, and sediment core (FLR1) used in this study. False River Lake is an oxbow lake that formed via neck cut-off of the lower Mississippi River in AD 1722²⁰ and is situated outside the modern floodway. Bathymetric contours

(white) given in metres. Shaded relief shows relative topographic lows (dark shades) and highs (light shades) according to the National Elevation Dataset²⁵.

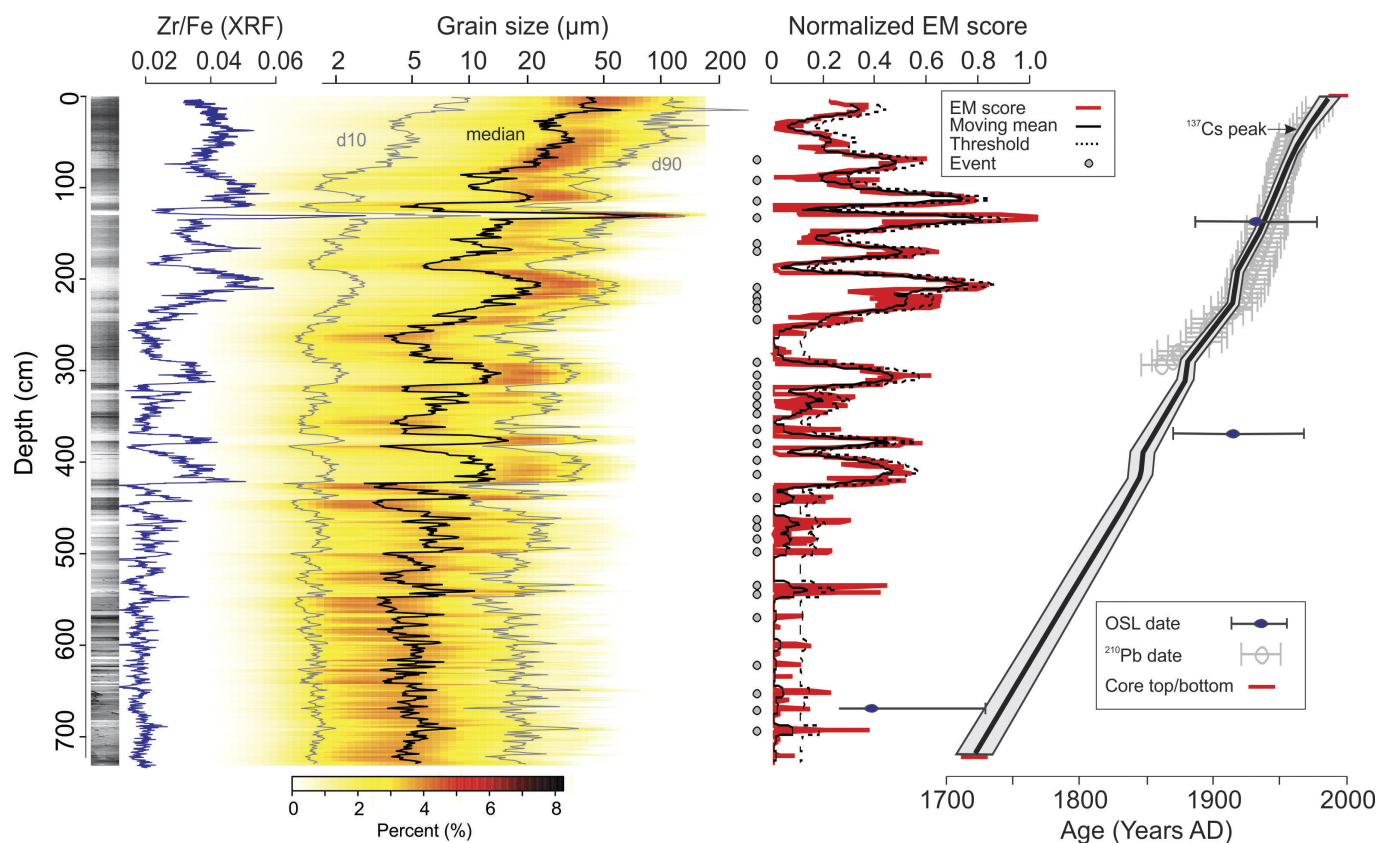


Extended Data Figure 3 | Location of Lake Saint John, Louisiana, and sediment core (STJ1) used in this study. Lake Saint John is an oxbow lake that formed via neck cut-off of the lower Mississippi River in about AD 1500²⁰ and is situated outside the modern floodway. Bathymetric contours

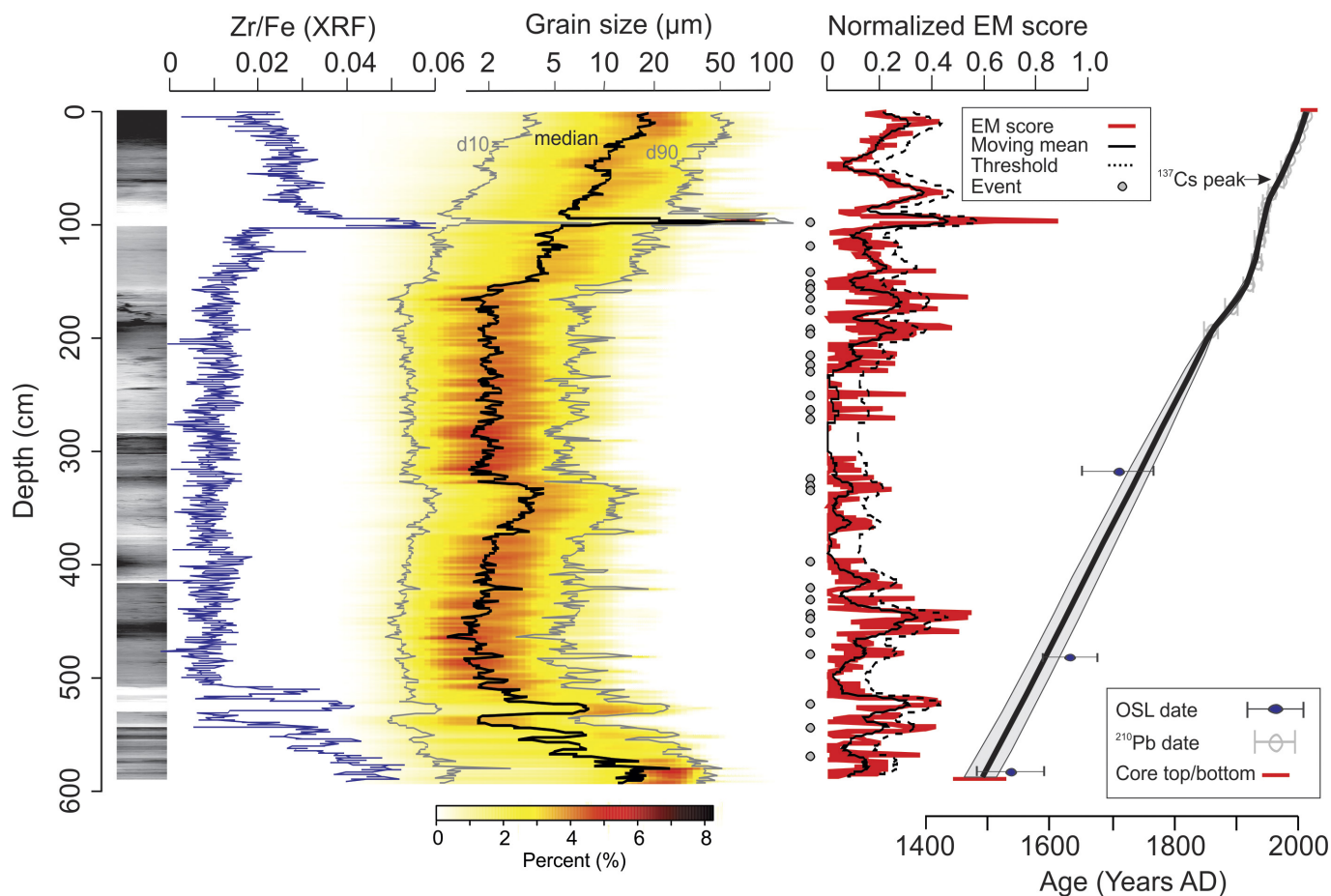
(white) given in metres. Shaded relief shows relative topographic lows (dark shades) and highs (light shades) according to the National Elevation Dataset²⁵.



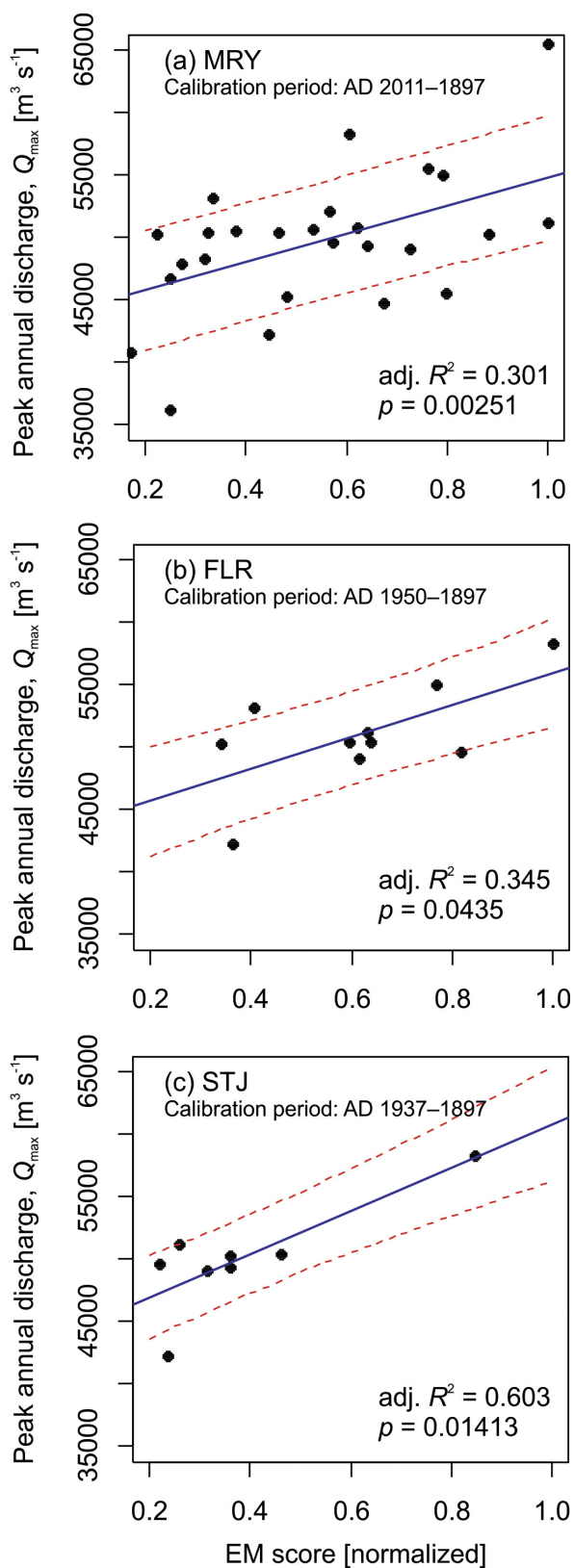
Extended Data Figure 4 | Radiography, bulk geochemistry, grain size and chronology of core MRY2. The age–depth model at right shows the median age probability (black line) and 1 σ confidence intervals (grey shading), with 2 σ confidence intervals on individual chronological controls.



Extended Data Figure 5 | Radiography, bulk geochemistry, grain size and chronology of core FLR1. The age–depth model at right shows the median age probability (black line) and 1 σ confidence intervals (grey shading), with 2 σ confidence intervals on individual chronological controls.



Extended Data Figure 6 | Radiography, bulk geochemistry, grain size and chronology of core STJ1. The age–depth model at right shows the median age probability (black line) and 1σ confidence intervals (grey shading), with 2σ confidence intervals on individual chronological controls.



Extended Data Figure 7 | Relationships between peak annual discharge and normalized EM score for historical floods in sedimentary archives.

Scatterplots and linear regressions with 1σ prediction intervals relating normalized EM score (a measure of grain size) to peak annual discharge of historical flood events for (a) MRY, (b) FLR and (c) STJ. Peak annual discharge estimates are from the Mississippi River gauging station at Vicksburg. Calibration periods vary owing to site-specific factors discussed in the Methods and Supplementary Information. adj., adjusted.

Shifts in tree functional composition amplify the response of forest biomass to climate

Tao Zhang¹, Ülo Niinemets², Justin Sheffield^{3,4} & Jeremy W. Lichstein¹

Forests have a key role in global ecosystems, hosting much of the world's terrestrial biodiversity and acting as a net sink for atmospheric carbon¹. These and other ecosystem services that are provided by forests may be sensitive to climate change as well as climate variability on shorter time scales (for example, annual to decadal)^{2–4}. Previous studies have documented responses of forest ecosystems to climate change and climate variability^{2–4}, including drought-induced increases in tree mortality rates⁵. However, relationships between forest biomass, tree species composition and climate variability have not been quantified across a large region using systematically sampled data. Here we use systematic forest inventories from the 1980s and 2000s across the eastern USA to show that forest biomass responds to decadal-scale changes in water deficit, and that this biomass response is amplified by concurrent changes in community-mean drought tolerance, a functionally important aspect of tree species composition. The amplification of the direct effects of water stress on biomass occurs because water stress tends to induce a shift in tree species composition towards species that are more tolerant to drought but are slower growing. These results demonstrate concurrent changes in forest species composition and biomass carbon storage across a large, systematically sampled region, and highlight the potential for climate-induced changes in forest ecosystems across the world, resulting from both direct effects of climate on forest biomass and indirect effects mediated by shifts in species composition.

How forests respond to climate variability has important implications for their future provisioning of ecosystem services, including carbon storage, timber, wildlife habitats and regulation of the hydrological cycle^{1,2}. Long-term demographic and geographic responses of tree species to climate change have been documented^{6,7}, and numerous studies have reported the effects of drought and other extreme events on tree growth and mortality^{2–5}. However, we have limited understanding of how the functional composition of tree communities responds to climate variability. It is not known how fast functional shifts occur; for example, whether one or two decades of relatively dry or wet conditions are sufficient to shift tree communities towards more or less drought-tolerant species. It is also unclear how such shifts affect the response of forest biomass (a key component of terrestrial carbon storage) to climate variability; that is, whether functional shifts moderate drought-induced biomass loss (as suggested by ecosystem model projections that demonstrate how such shifts can increase forest resilience to climate change⁸) or whether functional shifts amplify drought-induced biomass loss (for example, if competition under water-limited conditions drives plants to invest more in roots than is optimal for biomass production⁹).

The eastern USA has experienced substantial climate variability over recent decades, with some areas becoming wetter and others drier¹⁰ (Fig. 1a and Extended Data Fig. 1a–d). This variability—combined with systematically sampled Forest Inventory and Analysis¹¹ (FIA) data spanning millions of individual-tree records over several

decades—provides a valuable opportunity to quantify climate-induced changes in forest biomass and species composition. Forests of the eastern USA are influenced by various factors, including regrowth and successional dynamics after agricultural abandonment and logging¹², fire suppression¹³, introduced pathogens and insects¹⁴, increases in deer (*Odocoileus virginianus*) populations¹⁵ and natural disturbances¹⁶. Despite the importance of these non-climatic factors, correlations between precipitation change and recent shifts in the abundances and geographic ranges of tree species¹⁷ suggest that climate variability has a tangible effect on forest dynamics in the eastern USA, as observed elsewhere¹⁸. However, there is limited understanding of how shifts in forest functional composition affect the response of ecosystem properties, such as forest biomass, to climate variability.

To gain insights into how functional shifts affect the responses of ecosystems to climate variability, we quantified relationships among tree functional composition, forest biomass and water availability. We used the Palmer drought severity index (PDSI, an index of soil moisture based on the balance between precipitation and modelled evapotranspiration and run-off¹⁹) to quantify changes in mean growing season water availability (Δ PDSI) from the 1980s to 2000s in 1° latitude × 1° longitude grid cells (Fig. 1a). A one-unit decrease in growing-season PDSI is equivalent, on average, to a 23% reduction in growing-season precipitation in the eastern USA (Extended Data Fig. 1i–k). To quantify changes in community-mean drought tolerance (Δ DT, a functional component of tree species composition) and above-ground biomass (Δ AGB; Fig. 1 and Extended Data Figs 2, 3), we used FIA data while controlling for changes in forest stand age due to recovery from previous disturbance. Specifically, Δ DT and Δ AGB were calculated for each grid cell by comparing inventory plots in a given age class in the 1980s to plots in the same age class in the 2000s. Δ DT was calculated by combining FIA data in each decade, grid cell and age class with a species-specific drought-tolerance ranking (DT²⁰; Supplementary Methods 1), which increases from 1 (very intolerant) to 5 (very tolerant); therefore, Δ DT > 0 indicates a shift in tree species composition towards a more drought-tolerant community (see examples in Supplementary Methods 2). Although the DT scale is arbitrary, such rankings are widely used by plant ecologists and foresters^{10,17,21}, and modifying the scale would not qualitatively affect our results (Supplementary Methods 1). DT has been estimated for most of the common tree species in the eastern USA²⁰, enabling a systematic, community-level analysis of around 3,000,000 individual-tree records from around 100,000 FIA plot inventories (Extended Data Fig. 4a).

Relationships among Δ DT, Δ AGB and Δ PDSI were statistically significant ($P \leq 0.05$) in most of the 15 cases (3 pairwise correlations × 5 age classes). Specifically, from the 1980s to the 2000s, community-mean drought tolerance tended to increase with increasing water stress (negative correlation between Δ DT and Δ PDSI; Fig. 2a), biomass tended to increase with water availability (positive correlation between Δ AGB and Δ PDSI; Fig. 2b) and biomass tended to decrease with increasing community-mean drought tolerance (negative correlation

¹Department of Biology, University of Florida, Gainesville, Florida, USA. ²Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Tartu, Estonia. ³Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA. ⁴Geography and Environment, University of Southampton, Southampton, UK.

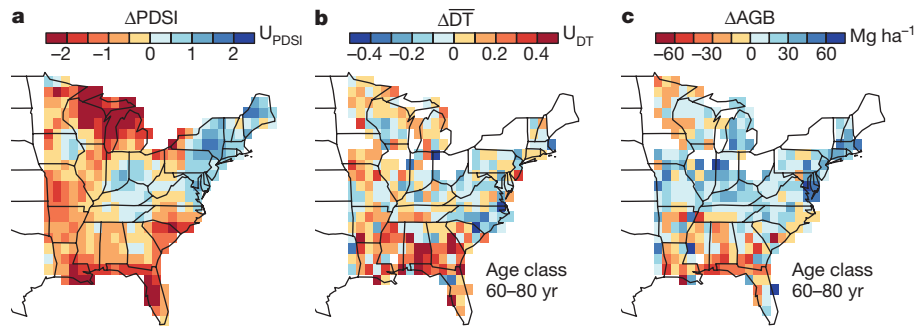


Figure 1 | Changes in growing-season Palmer drought severity index, community-mean drought tolerance and above-ground biomass between the 1980s and 2000s in the eastern USA. a, ΔPDSI . b, $\Delta\overline{\text{DT}}$. c, ΔAGB . $\Delta\overline{\text{DT}}$ and ΔAGB were calculated by comparing forest inventory plots that were 60–80 years old in the 1980s to plots in the same age class

in the 2000s (other age classes in Extended Data Figs 2, 3). Colour scales are oriented so that red corresponds to decreased moisture (in PDSI units; U_{PDSI}), increased $\overline{\text{DT}}$ (in DT units; U_{DT}) and decreased AGB. White grid cells lack sufficient inventory data.

between ΔAGB and $\Delta\overline{\text{DT}}$; Fig. 2c). These correlations suggest that there are concerted changes in forest functional composition and biomass in response to changes in water availability.

To determine whether the above correlations (Fig. 2) were robust, we performed additional analyses to quantify relationships among $\Delta\overline{\text{DT}}$, ΔAGB and ΔPDSI while controlling for changes in forest stand age and other potentially confounding variables (length of growing season, community-mean shade tolerance, tree harvesting, and/or spatial autocorrelation that may arise from factors not included in our statistical models; Extended Data Figs 5, 6 and Supplementary Methods 3). We used three statistical methods (spatial autoregressive models, structural equation modelling and independent effects analysis), all of which yielded consistent results (compare Extended Data Figs 4, 6 and 7) and demonstrate that the correlations (Fig. 2) are qualitatively robust. The analyses also revealed variation among forest age classes in terms of the relative importance of different variables affecting $\Delta\overline{\text{DT}}$ and ΔAGB (Extended Data Fig. 7), highlighting the need to study forest dynamics across successional stages and across life stages (for example, seedlings versus adults). However, in all age classes, $\Delta\overline{\text{DT}}$ had a strong negative correlation with changes in community-mean shade tolerance ($\Delta\overline{\text{ST}}$; Extended Data Fig. 7d), as expected from the interspecific trade-off between tolerances to shade and drought²⁰. Shifts in $\overline{\text{ST}}$ within forest age classes may reflect changing disturbance regimes¹³ or may be simply a consequence of PDSI-induced changes in $\overline{\text{DT}}$. Although our methods cannot determine whether changes in $\overline{\text{ST}}$ are a cause or a consequence of changes in $\overline{\text{DT}}$, our analyses demonstrate significant relationships among $\Delta\overline{\text{DT}}$, ΔAGB and ΔPDSI that are independent of $\Delta\overline{\text{ST}}$ and other potentially confounding variables. These results are further corroborated by a stand-level analysis that tracks the dynamics of individual FIA plots that have been measured and remeasured since

the 1990s (Supplementary Methods 4). This stand-level analysis shows that the qualitative patterns that emerge at the 1° grid-cell scale over two decades are also detectable at the stand level over shorter time periods (five-year mean remeasurement interval; Extended Data Fig. 4d, e).

The effects of PDSI on AGB in some grid cells were of a similar magnitude to the effects of other important global change drivers on AGB. For example, compared to the case in which there is no change in PDSI, a decrease of two PDSI units (which was exceeded in 11% of grid cells; Extended Data Fig. 1c, d) would cause a reduction in AGB of 8.6–14.3 Mg ha⁻¹ over two decades (based on the slopes in Fig. 2b), or 7–19% of the mean AGB in different forest age classes (Extended Data Fig. 3). This response of AGB (equivalent to 0.21–0.36 Mg C ha⁻¹ yr⁻¹) is similar to the estimated effects of nitrogen deposition on AGB in the USA²² (0.12–0.37 Mg C ha⁻¹ yr⁻¹), as well as the AGB component of the USA forest carbon sink²³ (0.30–0.46 and 0.37–0.56 Mg C ha⁻¹ yr⁻¹, for the 1990–1999 and 2000–2007 periods, respectively), which reflects the combined effects of several drivers (for example, forest regrowth, nitrogen deposition and CO₂ fertilization). Our analyses suggest that in the absence of climate variability, the eastern-USA carbon sink would have been even stronger over recent decades, because the mean ΔPDSI across the eastern USA was -0.61 (Extended Data Fig. 1d), which suggests a weakening of the eastern-USA forest carbon sink by approximately 0.07–0.11 Mg C ha⁻¹ yr⁻¹.

The response of AGB to PDSI includes not only direct effects of ΔPDSI (that is, changes in AGB that would occur in the absence of changes in community-mean drought tolerance, $\overline{\text{DT}}$), but also indirect effects of ΔPDSI (that is, changes in AGB caused by PDSI-induced changes in $\overline{\text{DT}}$; Extended Data Fig. 8a, b). Our analysis reveals that direct and indirect effects both work in the same direction, which

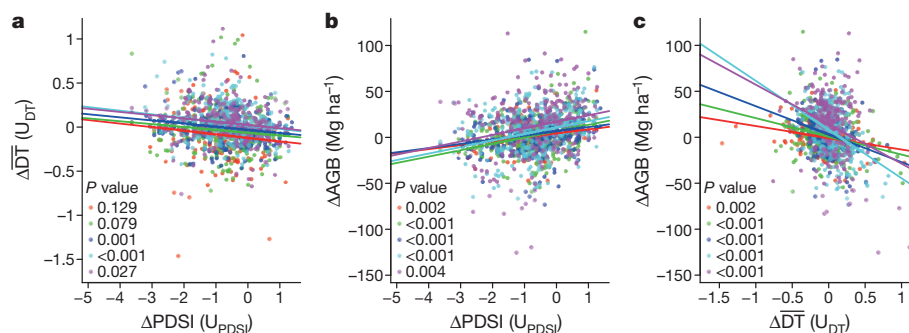


Figure 2 | Relationships between changes in community-mean drought tolerance, water availability and above-ground biomass within forest age classes from the 1980s to 2000s. a–c, Relationships between $\Delta\overline{\text{DT}}$ and ΔPDSI (a), ΔAGB and ΔPDSI (b) and ΔAGB and $\Delta\overline{\text{DT}}$ (c). Each point represents a forest age class within a 1° grid cell. Lines are ordinary least

squares regressions. *P* values (two-sided) for regression slopes are for the following age classes (from top to bottom): 0–20, 20–40, 40–60, 60–80 and 80–100 years. Pearson correlation ranges are 0.11–0.16 (a), 0.22–0.30 (b) and 0.24–0.40 (c). With outliers removed (not shown), all correlations have *P* < 0.05. Sample sizes are in Extended Data Fig. 4a.

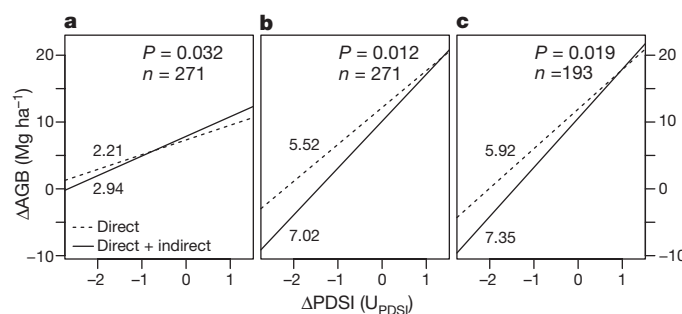


Figure 3 | Changes in species composition amplify the response of biomass to climate variability. **a–c**, ΔAGB versus ΔPDSI relationships for the three forest age classes with a significant response of $\Delta\overline{\text{DT}}$ to ΔPDSI : 40–60 years (**a**), 60–80 years (**b**) and 80–100 years (**c**). Dashed lines show direct effects of ΔPDSI on ΔAGB (controlling for $\Delta\overline{\text{DT}}$); solid

lines show total (direct + indirect) effects (including $\Delta\overline{\text{DT}}$; Extended Data Fig. 8). Lines are labelled with slopes ($\text{Mg ha}^{-1} \text{U}_{\text{PDSI}}^{-1}$). P values are based on s.e.m. of slope differences; n is number of grid cells. The x axes span the mean ± 2 s.d. from Fig. 1a.

means that changes in species composition (indirect effects) amplify the response of AGB to PDSI (Fig. 3). Indirect effects accounted for roughly 20–30% of the total response of ΔAGB to ΔPDSI in 40- to 100-year-old forests, significantly amplifying AGB responses in these age classes (Extended Data Fig. 8c, d).

The amplification of the biomass response to climate variability (Fig. 3) is probably driven by competitive shifts in the traits of the dominant tree species. For example, as water availability declines, competition may favour tree species that allocate more carbon to fine roots and less to leaves and wood than is optimal for biomass production⁹. This shift in allocation amplifies the response of the ecosystem to decreasing water availability, because the most competitive species under dry conditions have lower biomass than other species that could have persisted were they not outcompeted.

Shifts in community composition towards more- or less-drought-tolerant species may be caused by changes in relative species rankings in one or more demographic parameters (mortality, growth and recruitment), and are not necessarily caused by die-offs due to physiological stress. We decomposed stand-level $\Delta\overline{\text{DT}}$ into different components using data from remeasured inventory plots (Supplementary Methods 4, 5). Mortality, growth and recruitment all contributed substantially to $\Delta\overline{\text{DT}}$ (Fig. 4a), with the vast majority of $\Delta\overline{\text{DT}}$ accounted for by shifts in the abundance of species present at both measurement times (as opposed to species additions or losses; Fig. 4b). Mortality and species loss were most important in old stands, and recruitment and

species additions were most important in young stands (Fig. 4). These age-dependent trends probably reflect the larger size but lower density of trees in older forests; for example, the death of a single large tree can substantially affect $\overline{\text{DT}}$ (a size-weighted community average), whereas juvenile recruitment would have little immediate effect on $\overline{\text{DT}}$ in a mature forest.

Consistent with the importance of abundance shifts at the stand-level, grid-cell-scale relationships among $\Delta\overline{\text{DT}}$, ΔAGB and ΔPDSI reflect the collective responses of many species, without major changes in regional species diversity. Analysis of the influence of individual species on the changes in community-level $\overline{\text{DT}}$ and AGB revealed no systematic differences between angiosperms and gymnosperms (Extended Data Fig. 9), and no systematic relationships between species influence and either shade tolerance or wood density (Supplementary Methods 6, 7 and Supplementary Tables 1–5). Thus, the $\Delta\overline{\text{DT}}$ and ΔAGB responses reflect the collective responses of both angiosperm and gymnosperm species spanning a wide range of ecological strategies. Species abundance distributions and the number of common species within sub-regions (north-central, northeastern and southeastern USA) were similar between the 1980s and 2000s (Extended Data Fig. 10) despite substantial changes in the abundance of some individual species (Supplementary Tables 6–20). Some of these species-level changes were inconsistent with the overall community-level $\Delta\overline{\text{DT}}$ and ΔAGB responses, and may reflect long-term changes in disturbance regimes rather than climate responses¹³. For example, increases in abundance of two widespread maple species, *Acer rubrum* and *Acer saccharum*, are probably due to fire suppression and the resulting mesophication of eastern USA forests¹³, rather than a response to ΔPDSI . Thus, our study does not indicate that climate-driven functional shifts are the primary mode of change in eastern USA forests. Nevertheless, our results support recent evidence that climate-induced shifts in species abundances and geographic ranges have occurred over recent decades in the eastern USA¹⁷, and here we demonstrate for the first time that these shifts have important consequences at the ecosystem level.

The response of eastern USA forests to climate variability, evident from shifts in both forest biomass and species composition, is of global importance for several reasons. First, these shifts—quantified from systematic, regional-scale forest inventories—suggest that even greater changes may be underway in other regions where recent drought and climate change have been more severe than in the eastern USA; for example, in Amazonia and western North America, where drought-induced increases in tree mortality have already been reported^{4,24,25}. Second, projected increases in the frequency and severity of extreme weather events in many regions of the world²⁶—combined with the sensitivity of forest biomass and species composition to climate variability documented here and elsewhere²⁷—suggests potential changes in global forests over the next century that are both ecologically and

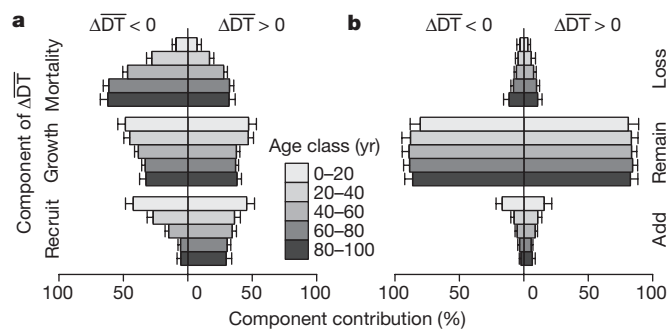


Figure 4 | Contributions of the different components to the change in community-mean drought tolerance. Component contributions are shown to the left of zero for negative $\Delta\overline{\text{DT}}$ (lower quartile of $\Delta\overline{\text{DT}}$ distribution), and to the right of zero for positive $\Delta\overline{\text{DT}}$ (upper quartile). Error bars show one side of each 95% confidence interval. **a**, Contributions of mortality, growth and recruitment to $\Delta\overline{\text{DT}}$. **b**, Contributions of species additions, losses and abundance shifts ('remain') to $\Delta\overline{\text{DT}}$. Sample sizes (number of remeasured plots within each quartile) are 300 (0–20 years), 588 (20–40 years), 1,173 (40–60 years), 1,390 (60–80 years) and 650 (80–100 years). See Supplementary Methods 5 for details.

economically important. Finally, amplification of the biomass–climate response due to shifts in species composition (temporal beta diversity²⁸) contrasts with evidence that local (alpha) diversity increases ecosystem stability²⁹, including increased resistance to climate extremes³⁰. These contrasting effects of alpha and beta diversity highlight the need to better understand how different components of biodiversity, including changes in species composition, affect ecosystem functioning at different spatial and temporal scales.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 March 2016; accepted 21 February 2018.

Published online 21 March 2018.

1. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320**, 1444–1449 (2008).
2. Anderegg, W. R. L., Kane, J. M. & Anderegg, L. D. L. Consequences of widespread tree mortality triggered by drought and temperature stress. *Nat. Clim. Chang.* **3**, 30–36 (2013).
3. Frank, D. *et al.* Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts. *Glob. Chang. Biol.* **21**, 2861–2880 (2015).
4. Clark, J. S. *et al.* The impacts of increasing drought on forest dynamics, structure, and biodiversity in the United States. *Glob. Chang. Biol.* **22**, 2329–2352 (2016).
5. Allen, C. D. *et al.* A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *For. Ecol. Manage.* **259**, 660–684 (2010).
6. Davis, M. B. & Shaw, R. G. Range shifts and adaptive responses to Quaternary climate change. *Science* **292**, 673–679 (2001).
7. Parmesan, C. & Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**, 37–42 (2003).
8. Sakschewski, B. *et al.* Resilience of Amazon forests emerges from plant trait diversity. *Nat. Clim. Chang.* **6**, 1032–1036 (2016).
9. Farrior, C. E., Dybzinski, R., Levin, S. A. & Pacala, S. W. Competition for water and light in closed-canopy forests: a tractable model of carbon allocation with implications for carbon sinks. *Am. Nat.* **181**, 314–330 (2013).
10. Peters, M. P., Iverson, L. R. & Matthews, S. N. Long-term droughtiness and drought tolerance of eastern US forests over five decades. *For. Ecol. Manage.* **345**, 56–64 (2015).
11. Bechtold, W. A. & Patterson, P. L. (eds.) *The Enhanced Forest Inventory and Analysis Program – National Sampling Design and Estimation Procedures*. Gen. Tech. Rep. SRS-80 (US Department of Agriculture, 2005).
12. Houghton, R. A., Hackler, J. L. & Lawrence, K. T. The U.S. carbon budget: contributions from land-use change. *Science* **285**, 574–578 (1999).
13. Nowacki, G. J. & Abrams, M. D. Is climate an important driver of post-European vegetation change in the Eastern United States? *Glob. Chang. Biol.* **21**, 314–334 (2015).
14. Hicke, J. A. *et al.* Effects of biotic disturbances on forest carbon cycling in the United States and Canada. *Glob. Chang. Biol.* **18**, 7–34 (2012).
15. Côté, S. D., Rooney, T. P., Tremblay, J.-P., Dussault, C. & Waller, D. M. Ecological impacts of deer overabundance. *Annu. Rev. Ecol. Evol. Syst.* **35**, 113–147 (2004).
16. Vanderwel, M. C., Coomes, D. A. & Purves, D. W. Quantifying variation in forest disturbance, and its effects on aboveground biomass dynamics, across the eastern United States. *Glob. Chang. Biol.* **19**, 1504–1517 (2013).
17. Fei, S. *et al.* Divergence of species responses to climate change. *Sci. Adv.* **3**, e1603055 (2017).
18. Ruiz-Benito, P. *et al.* Climate- and successional-related changes in functional composition of European forests are strongly driven by tree mortality. *Glob. Chang. Biol.* **23**, 4162–4176 (2017).
19. Sheffield, J., Wood, E. F. & Roderick, M. L. Little change in global drought over the past 60 years. *Nature* **491**, 435–438 (2012).
20. Niinemets, Ü. & Valladares, F. Tolerance to shade, drought, and waterlogging of temperate northern hemisphere trees and shrubs. *Ecol. Monogr.* **76**, 521–547 (2006).
21. Cornwell, W. K. & Grubb, P. J. Regional and local patterns in plant species richness with respect to resource availability. *Oikos* **100**, 417–428 (2003).
22. Pinder, R. W. *et al.* Climate change impacts of US reactive nitrogen. *Proc. Natl Acad. Sci. USA* **109**, 7671–7675 (2012).
23. Pan, Y. *et al.* A large and persistent carbon sink in the world's forests. *Science* **333**, 988–993 (2011).
24. Phillips, O. L. *et al.* Drought sensitivity of the Amazon rainforest. *Science* **323**, 1344–1347 (2009).
25. Peng, C. *et al.* A drought-induced pervasive increase in tree mortality across Canada's boreal forests. *Nat. Clim. Chang.* **1**, 467–471 (2011).
26. IPCC. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change* (eds Field, C. B. *et al.*) (Cambridge Univ. Press, 2012).
27. Hanewinkel, M., Cullmann, D. A., Schelhaas, M.-J., Nabuurs, G.-J. & Zimmermann, N. E. Climate change may cause severe loss in the economic value of European forest land. *Nat. Clim. Chang.* **3**, 203–207 (2013).
28. Dornelas, M. *et al.* Assemblage time series reveal biodiversity change but not systematic loss. *Science* **344**, 296–299 (2014).
29. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol. Monogr.* **75**, 3–35 (2005).
30. Isbell, F. *et al.* Biodiversity increases the resistance of ecosystem productivity to climate extremes. *Nature* **526**, 574–577 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements Funding was provided by US Department of Agriculture Forest Service agreements 11-JV-11242306-059 and 16-JV-11242306-050 to J.W.L., and by the European Regional Development Fund (Centre of Excellence EcolChange) and the Estonian Ministry of Science and Education (institutional grant IUT-8-3) to Ü.N.

Author Contributions T.Z. and J.W.L. designed the research. J.S. and Ü.N. provided data and advice. T.Z. performed the analysis. T.Z. and J.W.L. drafted the first version of the paper, and all authors contributed to subsequent versions of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to T.Z. (t.zhang05@outlook.com) or J.W.L. (jlichstein@ufl.edu).

Reviewer Information *Nature* thanks C. Schwalm and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Overview. We used simple correlations and several multivariate methods to quantify relationships among changes in community mean drought-tolerance ($\Delta\overline{DT}$), above-ground forest biomass (ΔAGB) and growing-season water availability (ΔPDSI) in the contiguous USA east of 95° longitude. Most analyses were executed at the 1° latitude \times 1° longitude grid-cell scale and controlled for changes in stand age by comparing forest inventory plots in the 1980s to plots in the same grid cell and age class in the 2000s. The following sections describe these grid-level analyses. Stand-level analyses that tracked the dynamics of remeasured inventory plots are described in Supplementary Methods 4, 5. In addition to controlling for changes in stand age, our multivariate analyses also controlled for changes in community-mean shade-tolerance ($\Delta\overline{ST}$, which could be confounded with $\Delta\overline{DT}$ owing to the negative correlation between shade- and drought-tolerance²⁰) and growing-season length (ΔGSL , which could be confounded with ΔPDSI , since both depend on temperature). We further controlled for potentially confounding effects of unmeasured, spatially structured variables using simultaneous autoregressive (SAR) models^{31,32}. To evaluate the robustness of the SAR model results, we used two additional multivariate methods: structural equation modelling³³ (SEM) and independent effects analysis³⁴ (IEA).

Quantifying decadal climate variability. We used the change in the PDSI during the growing season from the 1980s to 2000s to quantify decadal climate variability. The growing season was defined as the period in a given year between the last frost (minimum daily temperature below 0 °C) in the spring and the first frost in the fall. We identified the growing season in each grid cell each year using temperature information in the 1° spatial, three-hourly, global climate reanalysis dataset (<http://hydrology.princeton.edu/data.pgf.php>) from ref. 35. Mean growing-season PDSI was calculated for each grid cell in each year as the weighted mean of monthly PDSI, in which the weights are the number of days in a given month that were included in the growing season. For example, if the growing season in a given grid cell and year began on May 20 and ended on September 10, then the weights for May, June, July, August and September, respectively, would be 12, 30, 31, 31 and 10. Monthly PDSI in each grid cell was calculated as the median of 16 versions of a global (1° latitude \times 1° longitude) PDSI dataset (<http://hydrology.princeton.edu/data.pdsi.php>) from ref. 19. The 16 PDSI versions are derived from the 16 combinations of 4 global precipitation datasets, 2 PDSI models and 2 calibration methods (see ref. 19 for details). Preliminary analyses showed that our results were qualitatively robust to variation among the 16 PDSI versions.

After calculating the mean growing-season PDSI for each grid cell in each year, the PDSI change (ΔPDSI) between the 1980s and the 2000s was quantified as $\Delta\text{PDSI} = \text{PDSI}_{2000s} - \text{PDSI}_{1980s}$, in which PDSI_{1980s} and PDSI_{2000s} are the means of the growing season PDSI during the two decades (1980–1989 and 2000–2009, respectively). Similarly, the change in growing season length (ΔGSL) between the two decades was quantified as $\Delta\text{GSL} = \text{GSL}_{2000s} - \text{GSL}_{1980s}$, in which GSL_{1980s} and GSL_{2000s} are the GSL means in the 1980s and 2000s, respectively.

Description of the US Forest Inventory and Analysis database. Tree species composition and biomass in each grid cell were quantified from tree-level data reported in the US Forest Inventory Analysis (FIA) database (<http://www.fia.fs.fed.us/>), which reports the diameter at breast height (DBH) and species identity of individual trees in forest inventory plots that are geographically distributed to provide unbiased estimates of forest biomass and tree population attributes¹¹. There is roughly one plot per 25 km², and each plot samples an area of approximately 0.1 ha. Since FIA's national standardized design was implemented (nominally in the year 1999, although the exact year varies by US state), plots in the eastern USA have been remeasured roughly every 5 years as follows: trees with DBH ≥ 12.7 cm are inventoried on four 7.32-m radius subplots per plot, and trees with DBH between 2.54 and 12.7 cm are inventoried on four 2.07-m radius microplots per plot¹¹. Prior to around 1999, sampling designs and remeasurement intervals varied among US states, but all state inventories aimed to yield unbiased population-level statistics¹¹. Most eastern USA states were inventoried in both the 1980s and 2000s, although the same plot locations were not always resampled. Some states were also inventoried in the 1990s, but we focused our analysis on changes from the 1980s to 2000s owing to inconsistent data availability for the 1990s. We assigned each FIA plot to a 1° grid cell based on approximate plot coordinates reported by FIA. Errors in plot coordinates reported in the public FIA database (see details in ref. 36) should have minimal effect on our 1° grid-cell-scale analyses. Owing to frequent discontinuities in sampling designs and/or plot locations, it is often not possible to track the dynamics of individual plots between the pre- and post-1999 periods. Furthermore, the unique plot identifiers needed to connect plot records across these periods are not provided in the public FIA dataset. However, it is possible to track the dynamics of individual plots during the post-1999 period, and we used plots remeasured during this period to perform stand-level analyses (Supplementary Methods 4).

All analyses presented in our paper are based on FIA plots in naturally regenerated (non-plantation) forest. However, many non-plantation forests in the USA are managed for timber or other wood products, with management regimes ranging from infrequent selective harvest to periodic clear-cutting. Our analyses account for stand-replacing disturbance (including intensive logging) by separately analysing different stand-age classes. Also, unless stated otherwise, our analyses minimized the effects of selective harvest by excluding data from FIA plot inventories where one or more harvested trees were reported. However, including selectively harvested plots and accounting for harvest effects does not qualitatively affect our results (Supplementary Methods 3 and Extended Data Fig. 6q).

Stand-age classes. To control for age-related changes in tree species composition and biomass, we controlled for changes in stand age in our grid-cell-scale analyses by comparing forest inventory plots in the 1980s to plots in the same grid cell and age class in the 2000s. We assigned each plot inventory to one of the following age classes based on stand ages reported by FIA (defined as the mean age of trees in the dominant size class³⁷): 0–20, 20–40, 40–60, 60–80 or 80–100 years. The width of each age class (20 years) matches the timescale of the change analysis (1980s to 2000s). Therefore, plot locations that were measured in both decades were typically assigned to different age classes in the 1980s and 2000s; for example, a plot that was in the 40–60 year age-class in the 1980s was typically in the 60–80 year age-class in the 2000s. Thus, our grid-cell-scale analysis of change from the 1980s to 2000s involves the comparison of two largely independent samples (that is, two different sets of plots with similar mean age), which minimizes the potential for successional dynamics to influence the results. Grid-cell-scale analyses for a given age class included only those grid cells for which at least five FIA plots in both the 1980s and 2000s met our filtering criteria; for example, plots without recent selective harvest and where species with unknown DT comprised <20% of plot AGB (see 'Quantifying decadal changes in forest tree species composition').

Quantifying decadal changes in forest tree species composition. For each age class within each grid cell and decade (1980s and 2000s), we quantified community-mean drought tolerance (\overline{DT}) based on the size and species identity of live trees in FIA plots as $\overline{DT} = \frac{\sum_{i=1}^n \text{AGB}_i \text{DT}_i}{\sum_{i=1}^n \text{AGB}_i}$, in which n is the number of individual trees; DT_i is the drought tolerance index for the species of individual i (the DT index is described in detail and evaluated with independent data in Supplementary Methods 1); and AGB_i is the estimated above-ground biomass density (Mg ha^{-1}) of individual i . To estimate AGB_i , we combined DBH measurements reported by FIA with biomass allometries derived for USA tree species groups³⁸ to estimate the above-ground biomass (Mg) of tree i , which was then converted to Mg ha^{-1} units by multiplying biomass (Mg) by the number of trees per hectare (ha^{-1}) represented by tree i . These ha^{-1} values are reported by FIA as trees per acre, which is the inverse of the sample area for tree i ; for example, in the national standardized plot design¹¹, the sample area for a tree with DBH ≥ 12.7 cm is the number of subplots per plot (four) \times the subplot area (0.0168 ha). For each grid cell, the change in community-mean drought tolerance was then calculated as $\Delta\overline{DT} = \overline{DT}_{2000s} - \overline{DT}_{1980s}$.

The DT index was unavailable for some eastern USA tree species (which accounted for 0.10% of the individuals in our analysis), and some individuals were identified only to the genus level in the FIA database (2.1% of individuals in our analysis). Individuals identified only to the genus level were split into expected contributions of different species as follows: if any congeneric individuals present in the same grid cell and age class (in either the 1980s or 2000s) were identified to species and belonged to a species with known DT, the 'trees per hectare' value (see above) of the genus-level individual was divided into different known-DT species in proportion to their relative abundances (based on AGB) in a given grid cell and age class (in the 1980s and 2000s combined). The above algorithm allowed us to assign DT values for most individuals with missing DT. We restricted our subsequent analyses to FIA plots for which individuals with unassigned DT (after splitting genus-level identifications to known-DT species as explained above) comprised less than 20% of plot AGB. The excluded plots accounted for only 0.24% and 0.16% of total plots in the 1980s and 2000s, respectively.

Similarly, we calculated community-mean shade tolerance (\overline{ST}) and its change ($\Delta\overline{ST}$) based on a species-specific shade-tolerance index (ST) from ref. 20, ranging from 1 (very intolerant) to 5 (very tolerant). As noted above, DT was unavailable for some species, and ST was unavailable for these same species. For individuals for which we replaced missing DT with assigned values, we also replaced missing ST using the same procedure as described above. We then calculated \overline{ST} for each grid cell, age class and decade as $\overline{ST} = \frac{\sum_{i=1}^n \text{AGB}_i \text{ST}_i}{\sum_{i=1}^n \text{AGB}_i}$; and we

calculated its change as $\Delta\overline{ST} = \overline{ST}_{2000s} - \overline{ST}_{1980s}$.

To evaluate whether our main results depended on using AGB weights to calculate \overline{DT} and \overline{ST} , we repeated our analyses using basal-area weights (proportional to tree diameter squared), which yielded very similar results.

Quantifying decadal changes in forest biomass. We estimated AGB (Mg ha^{-1}) for individual trees as described in the previous section. We then summed these individual AGB values to estimate plot-level AGB (Mg ha^{-1}) for each FIA plot, and we calculated average AGB across plots within each grid cell, decade and age class. The change in AGB density from the 1980s to 2000s was then calculated for each age class within each grid cell as $\Delta\text{AGB} = \text{AGB}_{2000\text{s}} - \text{AGB}_{1980\text{s}}$.

Quantifying the response of community-mean drought tolerance and biomass to climate variability using spatial regression models. We implemented SAR models for each age class (see subsection 'Stand-age classes') with the `spatolm` function in the 'spdep' package³² in R (<https://www.r-project.org>). We used the following SAR model to quantify the response of community-mean drought tolerance ($\Delta\overline{\text{DT}}$) to climate variability (ΔPDSI) while controlling for changes in growing season length (ΔGSL) and community-mean shade-tolerance ($\Delta\overline{\text{ST}}$):

$$\Delta\overline{\text{DT}} = d_0 \mathbf{1} + d_P \Delta\text{PDSI} + d_G \Delta\text{GSL} + d_S \Delta\overline{\text{ST}} + \mathbf{u} \quad (1)$$

in which bold terms are vectors of length n (the number of grid cells included in the analysis; $\mathbf{1}$ is a vector of 1s of length n); d_0 is the intercept; d_P , d_G , and d_S are the slopes associated with the corresponding predictors; and \mathbf{u} is a vector of spatially autocorrelated errors. This vector is defined as $\mathbf{u} = \lambda W\mathbf{u} + \boldsymbol{\varepsilon}$, in which W is the row-standardized $n \times n$ spatial weights matrix (based on inverse distances to the 8 neighbour grid cells); $\boldsymbol{\varepsilon}$ are spatially independent errors; and λ is the spatial autoregressive coefficient.

We used an analogous SAR model to quantify the response of above-ground biomass (ΔAGB) to climate variability (ΔPDSI) and shifts in community-mean drought tolerance ($\Delta\overline{\text{DT}}$):

$$\Delta\text{AGB} = a_0 \mathbf{1} + a_P \Delta\text{PDSI} + a_D \Delta\overline{\text{DT}} + a_G \Delta\text{GSL} + a_S \Delta\overline{\text{ST}} + \mathbf{u} \quad (2)$$

in which a_0 is the intercept; a_P , a_D , a_G and a_S are the slopes associated with the corresponding predictors; and \mathbf{u} is a vector of spatially autocorrelated errors as described above.

We assumed linear forms for regression models because Pearson linear and Spearman rank correlations yielded very similar results for bivariate relationships among $\Delta\overline{\text{DT}}$, ΔAGB and ΔPDSI (Fig. 2), and because partial regression plots from multiple regression models did not reveal strong or consistent nonlinearities. **Partitioning variation in $\Delta\overline{\text{DT}}$ and ΔAGB using structural equation modelling and independent effects analysis.** To evaluate the robustness of our SAR model results (that is, the effects of ΔPDSI on $\Delta\overline{\text{DT}}$, and the effects of ΔPDSI and $\Delta\overline{\text{DT}}$ on ΔAGB), we used two additional methods—SEM³³ and IEA³⁴—to quantify relationships between response and predictor variables.

SEM is based on networks of hypothesized causal relationships among variables and is designed to disentangle potential causal pathways in multivariate datasets³³. The SEM model structures we evaluated are illustrated in Extended Data Fig. 7a–c. We implemented SEM with the `sem` function in the 'lavaan' package³⁹ in R.

IEA compares models fit with all possible contributions of predictor variables to quantify the contribution of each variable to goodness-of-fit³⁴. For example, if the improvement of goodness-of-fit due to including a given variable x is always greater than the improvement due to including any other variable, then variable x would be identified by IEA as having the greatest contribution to explaining variance in the response variable. IEA was applied to non-spatial forms of the regression models described in the previous subsection using the `hier.part` function in the 'hier.part' package in R.

Quantifying direct and indirect effects of ΔPDSI on ΔAGB and the amplifying effect of $\Delta\overline{\text{DT}}$. We used both SAR and SEM approach to quantify 'direct' effects of ΔPDSI on ΔAGB (for example, due to tree-level physiological processes) versus 'indirect' effects that occur because of changes in species composition. Direct and indirect effects are illustrated in Extended Data Fig. 8a, b. To quantify these effects in SAR analysis, we combined equations (1) and (2):

$$\begin{aligned} \Delta\text{AGB} &= a_0 \mathbf{1} + a_P \Delta\text{PDSI} + a_D \Delta\overline{\text{DT}} \\ &= a_0 \mathbf{1} + a_P \Delta\text{PDSI} + a_D(d_0 \mathbf{1} + d_P \Delta\text{PDSI}) \end{aligned}$$

$$\text{Therefore: } \Delta\text{AGB} = (a_0 + a_D \cdot d_0) \mathbf{1} + (a_P + a_D \cdot d_P) \Delta\text{PDSI} \quad (3)$$

Thus, the total effect of ΔPDSI on ΔAGB is $a_P + a_D \cdot d_P$, the direct effect is a_P (which is simply the partial regression coefficient for the effect of ΔPDSI on ΔAGB), and the indirect effect is $a_D \cdot d_P$; that is, the effect of ΔPDSI on ΔAGB due to the combined effects of $\Delta\overline{\text{DT}}$ on ΔAGB (a_D) and ΔPDSI on $\Delta\overline{\text{DT}}$ (d_P). Terms for growing season length (ΔGSL) and community-mean shade tolerance

($\Delta\overline{\text{ST}}$) are omitted from equation (3) because we control for these sources of variation here by setting $\Delta\text{GSL} = \Delta\overline{\text{ST}} = 0$ to isolate the effects of ΔPDSI and $\Delta\overline{\text{DT}}$. We also omit the error terms, as we rely on the previously fit SAR models (equations (1) and (2)) for our analysis of equation (3).

The estimated effect of $\Delta\overline{\text{DT}}$ on the slope of ΔAGB versus ΔPDSI is amplifying (rather than moderating) because estimates for a_D and d_P are both negative (Extended Data Fig. 8c). Thus, their product is positive, which increases the steepness of the ΔAGB versus ΔPDSI slope (equation (3)). To formally test the hypothesis that $a_D \cdot d_P$ is greater than zero, we used two different approaches, which both confirmed that $a_D \cdot d_P$ is significantly greater than zero. The first approach tested the hypothesis (that is, $a_D \cdot d_P > 0$) analytically by calculating the estimates

$$\text{and standard errors of } a_D \cdot d_P, \text{ which are } \hat{a}_D \cdot \hat{d}_P \text{ and } \hat{a}_D \cdot \hat{d}_P \sqrt{\left(\frac{\sigma_{a_D}}{\hat{a}_D}\right)^2 + \left(\frac{\sigma_{d_P}}{\hat{d}_P}\right)^2},$$

respectively. Estimates of a_D and their standard errors (\hat{a}_D and σ_{a_D}) are available from the SAR model for ΔAGB (equation (2)); and estimates of d_P and their standard errors (\hat{d}_P and σ_{d_P}) are available from the SAR model for $\Delta\overline{\text{DT}}$ (equation (1)). P values from this first approach are reported in Fig. 3. The second approach yielded very similar P values and involved error propagation using Monte Carlo simulations, in which we sampled from the approximately multivariate-normal parameter distributions from the two separate models (equations (1) and (2)) to generate a distribution for the product $a_D \cdot d_P$.

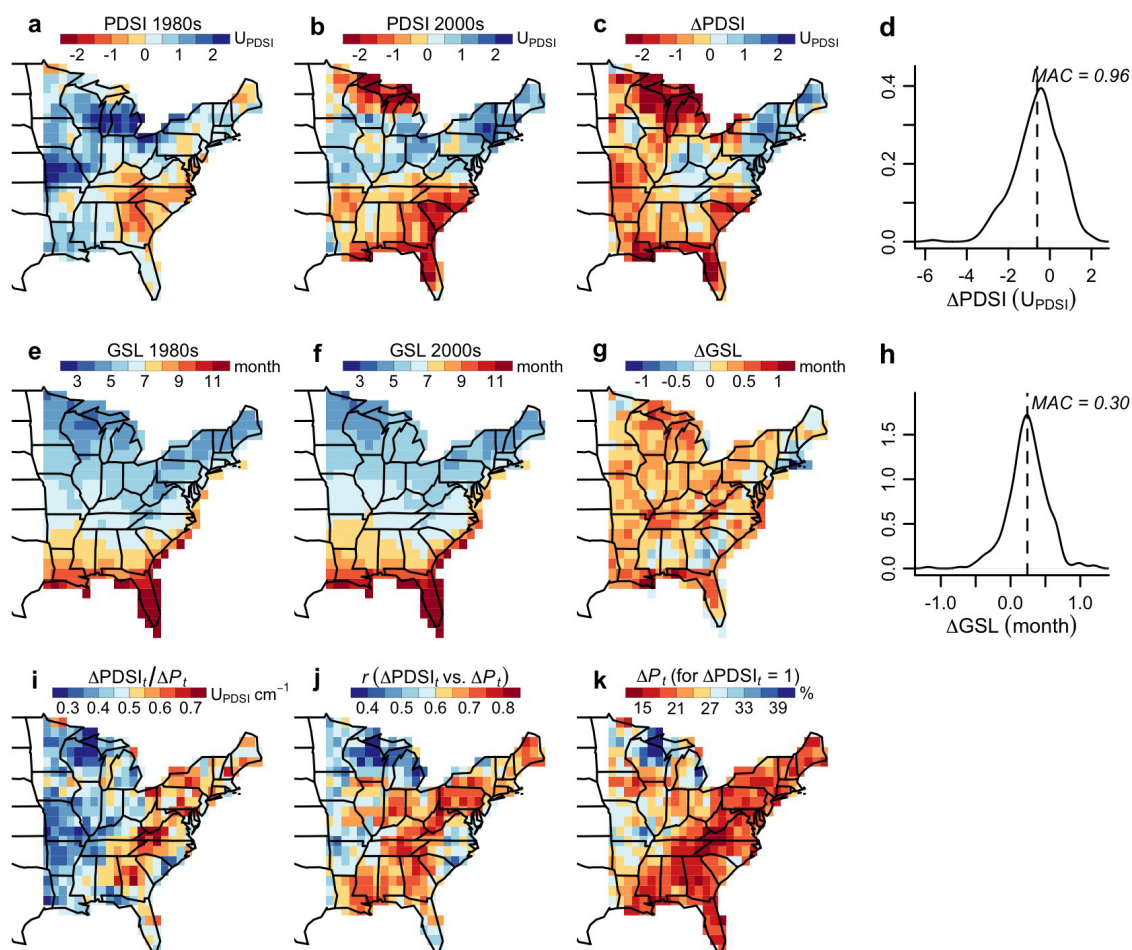
Comparison of PDSI effects on AGB with other global change drivers. We compared AGB responses in our study (based on the range of slopes in Fig. 2b) to the effects of nitrogen deposition²² and the mean USA forest carbon sink²³. All effects were expressed in AGB units of $\text{Mg C ha}^{-1} \text{ yr}^{-1}$, assuming a 2:1 AGB:C ratio.

Over recent decades, nitrogen deposition has increased carbon uptake in USA forests by an estimated $24\text{--}65 \text{ kg C kg N}^{-1}$, which includes 5 kg C kg N^{-1} in soil carbon and a multiplier of 1.2 to account for root biomass²². Removing the soil and root components yields the range $(24\text{--}5)/1.2$ to $(65\text{--}5)/1.2$, or $15.8\text{--}50 \text{ kg C kg N}^{-1}$. Estimates of USA forest area ($3.1 \times 10^8 \text{ ha}$) and nitrogen deposition (2.3 Tg N yr^{-1}) from ref. 22 yield a deposition rate of $7.42 \text{ kg N ha}^{-1} \text{ yr}^{-1}$. Multiplying this rate by the effect ($15.8\text{--}50 \text{ kg C kg N}^{-1}$) yields an estimated AGB increase of $0.12\text{--}0.37 \text{ Mg C ha}^{-1} \text{ yr}^{-1}$.

The USA forest carbon sink estimates are $0.72 \text{ Mg C ha}^{-1} \text{ yr}^{-1}$ (for the period 1990–1999; 66% in biomass) and $0.94 \text{ Mg C ha}^{-1} \text{ yr}^{-1}$ (for the period 2000–2007 period; 62% in biomass), based on table 2 of ref. 23. Assuming that 80% of biomass is AGB³⁸ yields $0.72 \times 66\% \times 80\%$ (1990–1999) and $0.94 \times 62\% \times 80\%$ (2000–2007) $\text{Mg C ha}^{-1} \text{ yr}^{-1}$. Applying $\pm 20\%$ uncertainty (which is the estimated 95% confidence range for USA forests in ref. 23) to these estimates yields 95% confidence intervals of $0.30\text{--}0.46$ (1990–1999) and $0.37\text{--}0.56$ (2000–2007) $\text{Mg C ha}^{-1} \text{ yr}^{-1}$ for the AGB component of the USA forest carbon sink.

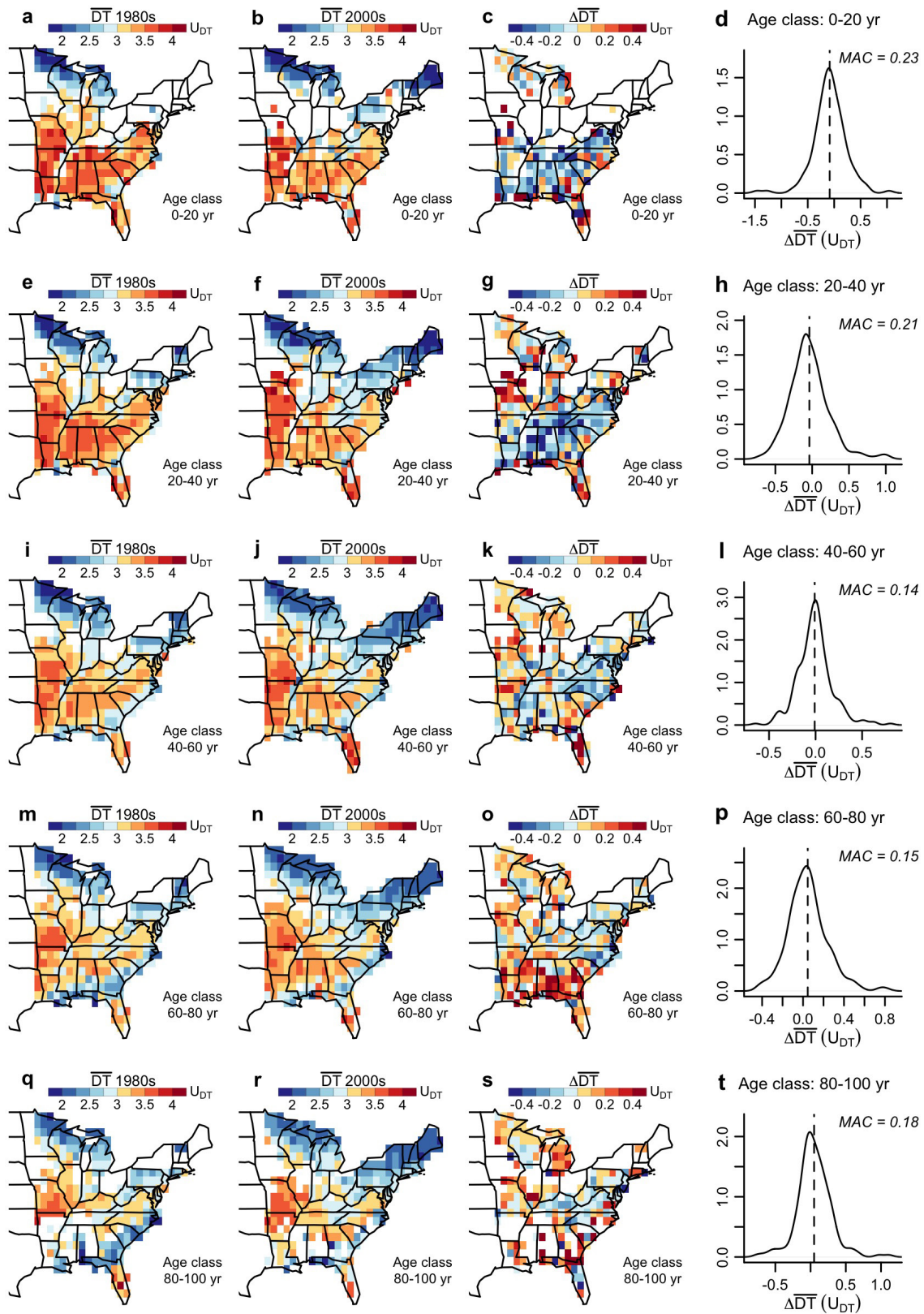
Data availability. All data used in our analyses are publicly available. FIA forest inventory data are available at <https://www.fia.fs.fed.us/>. Temperature and precipitation data are available at <http://hydrology.princeton.edu/data.pgf.php>. PDSI data are available at <http://hydrology.princeton.edu/data.pdsi.php>. The originally published versions of the DT and ST indices are available at <http://www.esapubs.org/Archive/mono/M076/020/appendix-A.htm>, and updates are described in Supplementary Methods 1. The US Census Bureau is the source of the reference map (US state boundaries).

- Lichstein, J. W., Simons, T. R., Shiner, S. A. & Franzreb, K. E. Spatial autocorrelation and autoregressive models in ecology. *Ecol. Monogr.* **72**, 445–463 (2002).
- Bivand, R. S., Hauke, J. & Kossowski, T. Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods. *Geogr. Anal.* **45**, 150–179 (2013).
- Grace, J. B. *Structural Equation Modeling and Natural Systems* (Cambridge Univ. Press, 2006).
- Chevan, A. & Sutherland, M. Hierarchical partitioning. *Am. Stat.* **45**, 90–96 (1991).
- Sheffield, J., Goteti, G. & Wood, E. F. Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling. *J. Clim.* **19**, 3088–3111 (2006).
- O'Connell, B. M. et al. *The Forest Inventory and Analysis Database: Database Description and User Guide for Phase 2 v.5.1.5* (US Department of Agriculture, 2013).
- Forest Inventory and Analysis National Core Field Guide. Volume 1: Field Data Collection Procedures for Phase 2 Plots, v.6.0* (US Department of Agriculture, 2012).
- Jenkins, J. C., Chojnacki, D. C., Heath, L. S. & Birdsey, R. A. National-scale biomass estimators for United States tree species. *For. Sci.* **49**, 12–35 (2003).
- Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, (2012).



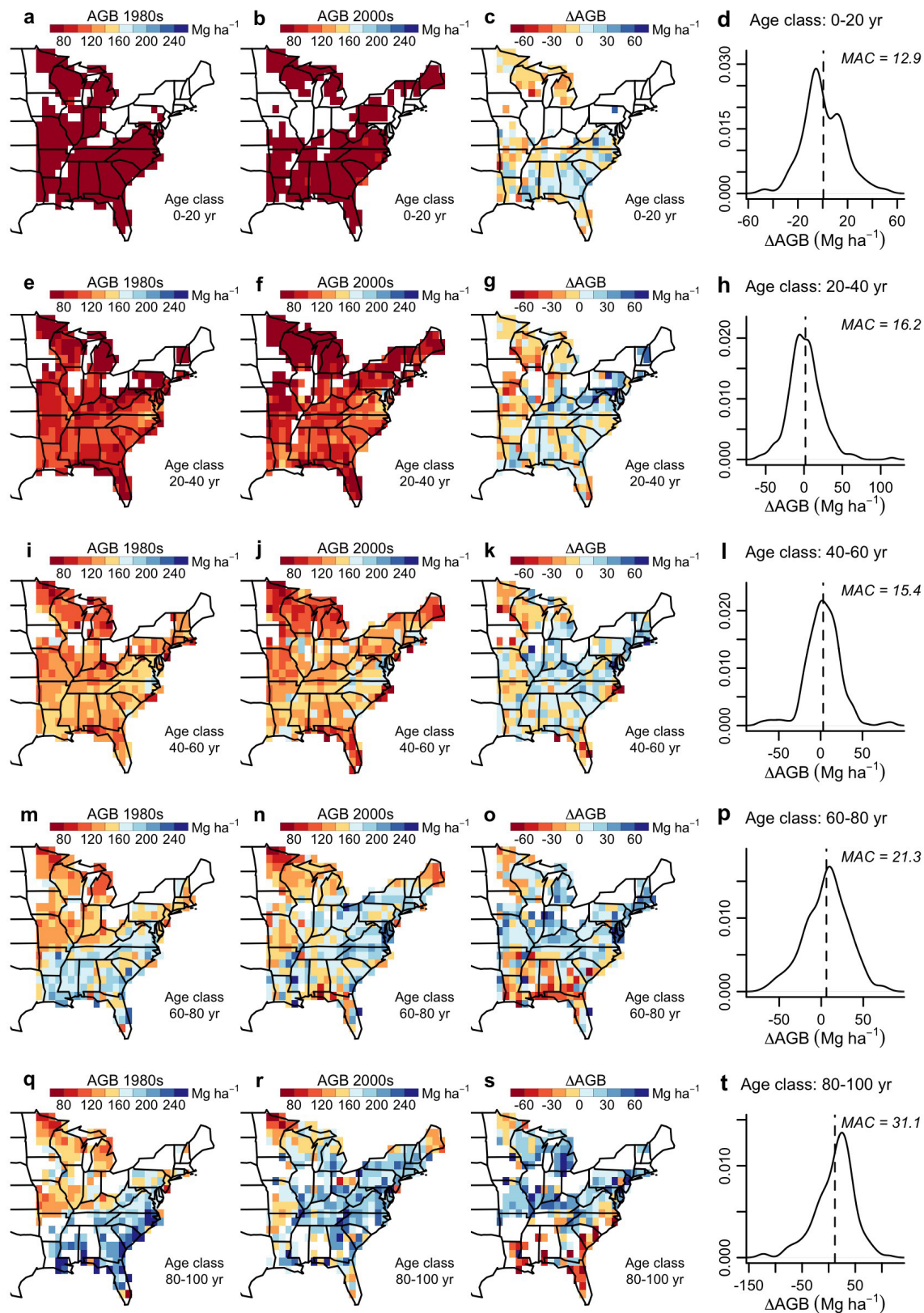
Extended Data Figure 1 | PDSI and growing-season length in the 1980s and 2000s, and relationships between PDSI and precipitation in the eastern USA. Maps show 1° latitude \times 1° longitude grid cells. **a–c**, Mean growing-season PDSI (see Methods) for the 1980s (**a**), the 2000s (**b**) and the change in PDSI (Δ PDSI) in PDSI units (U_{PDSI}) (**c**) between the two decades. **d**, Smoothed distribution of Δ PDSI; the vertical line is the mean, with the mean absolute change (MAC; mean of absolute values) indicated. **e–g**, Mean GSL for the 1980s (**e**), the 2000s (**f**) and the change in GSL (Δ GSL, months) (**g**) between the two decades. **h**, Smoothed distribution of Δ GSL; the vertical line is the mean, with the MAC indicated. **i**, The response of PDSI to precipitation change. The map shows regression slopes

of Δ PDSI_{*t*} (change in growing-season PDSI between successive years from 1948 to 2009) versus Δ P_{*t*} (change in growing-season precipitation between successive years); slope units are PDSI units (U_{PDSI}) per cm precipitation. The sample size for each regression (one regression per grid cell) is 61 (the number of annual changes from 1948–2009). **j**, Pearson's correlation (*r*) between Δ PDSI_{*t*} and Δ P_{*t*}; samples sizes as in **i**. **k**, The percentage change in growing-season precipitation corresponding to a one-unit change in PDSI; percentages were calculated as 100 times the inverse of the slope from **i** divided by the mean (1948–2009) growing-season precipitation for each grid cell.



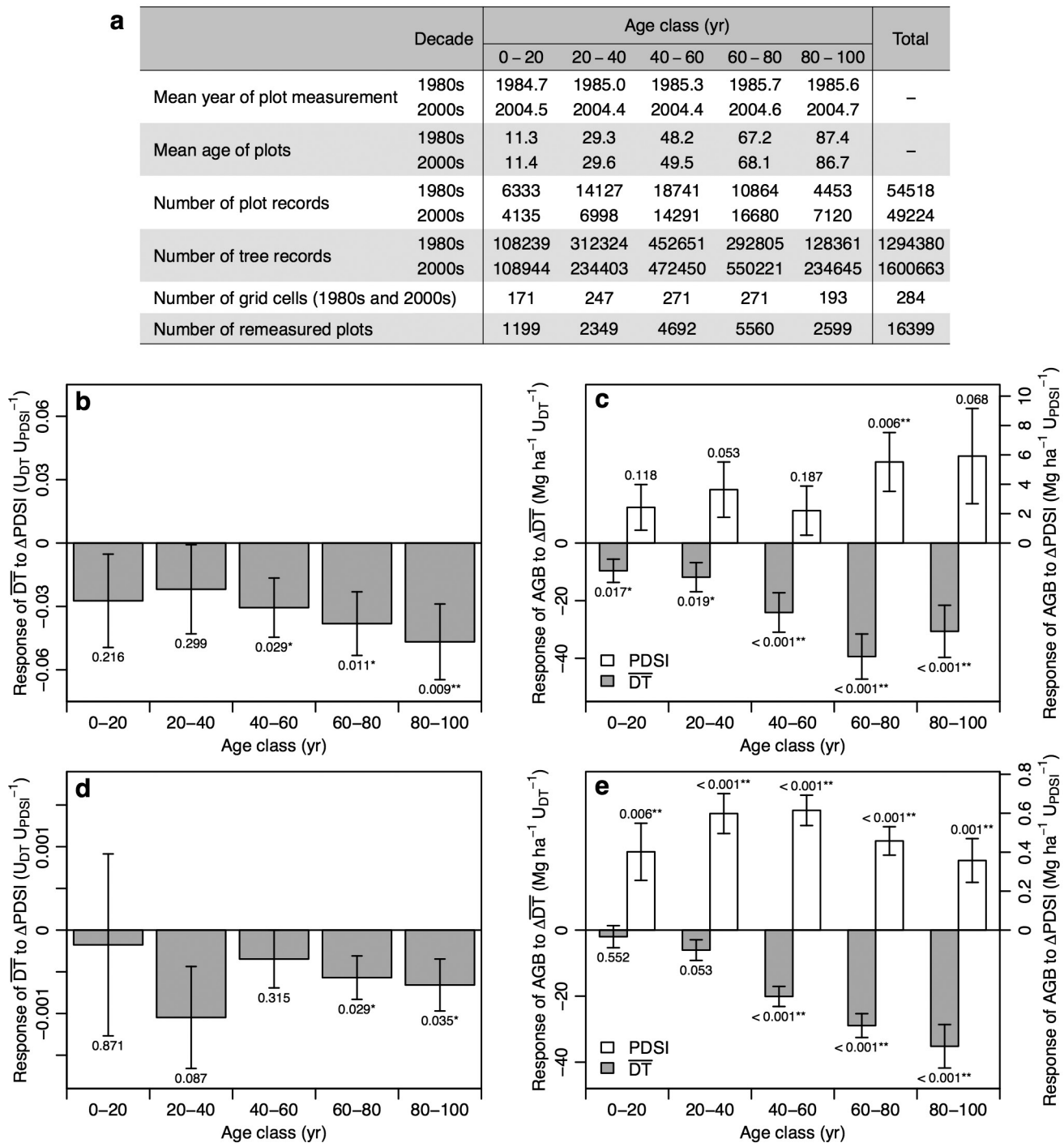
Extended Data Figure 2 | Community-mean drought tolerance in the 1980s and 2000s in the eastern USA. \overline{DT} was calculated within 1° grid cells and 20-year stand-age classes that contained at least five FIA inventory plots that satisfied filtering criteria (see Methods). From left, the columns show values of \overline{DT} for the 1980s, the 2000s, the change between the two decades and the smoothed distributions of changes (vertical lines

show mean changes, with MAC indicated). Different stand-age classes are shown in each row as follows: **a–d**, 0–20 years; **e–h**, 20–40 years; **i–l**, 40–60 years; **m–p**, 60–80 years; **q–t**, 80–100 years. \overline{DT} units (U_{DT}) are on the scale of the species drought tolerance (DT) index, which increases from 1 (very intolerant) to 5 (very tolerant); see Methods and Supplementary Methods 1 for details.



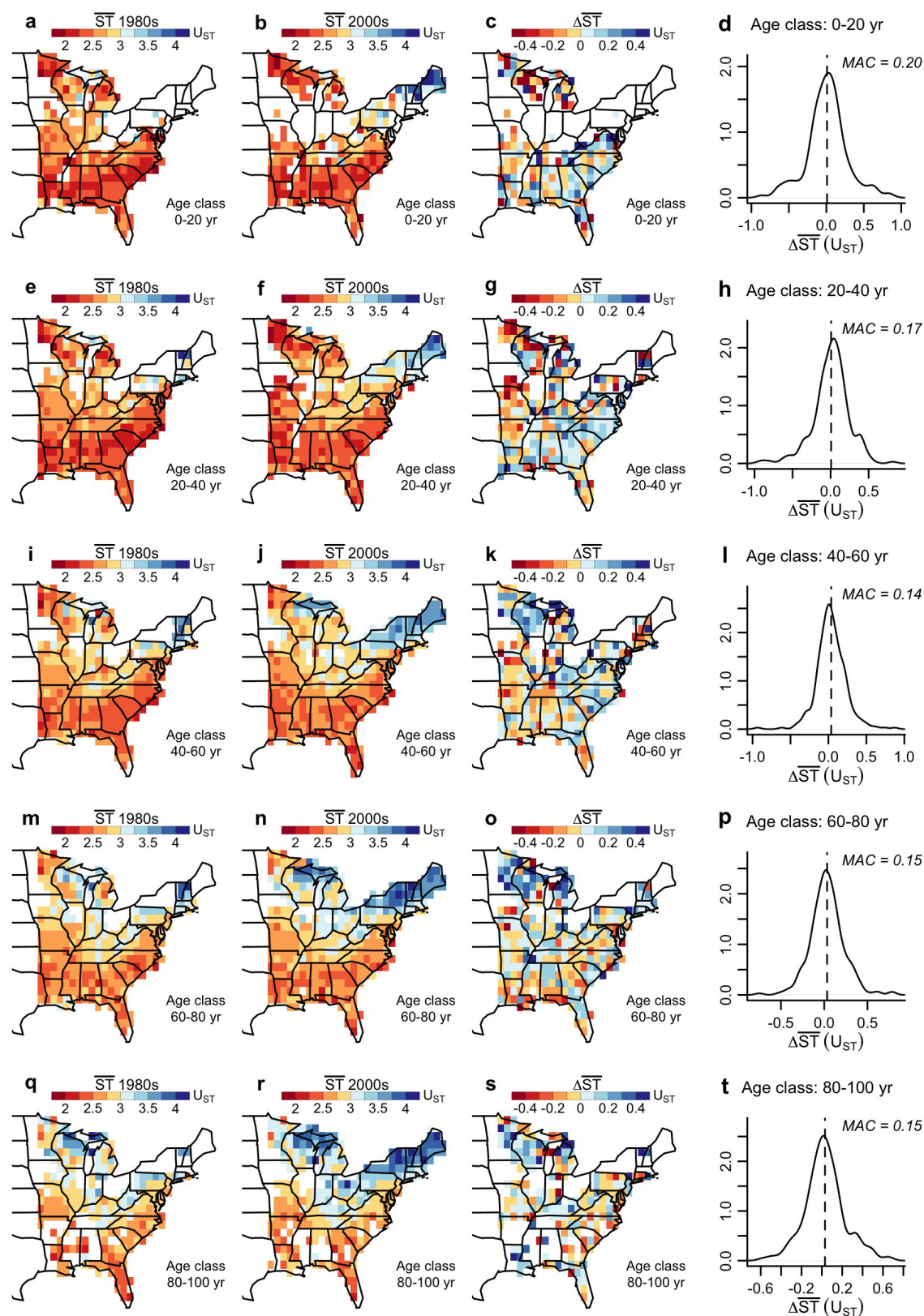
Extended Data Figure 3 | Above-ground live biomass in the 1980s and 2000s in the eastern USA. AGB (Mg ha^{-1}) was estimated from allometries and averaged within 1° grid cells and 20-year stand-age classes that contained at least five FIA inventory plots that satisfied filtering criteria (see Methods). From left, the columns show values of AGB for the

1980s, the 2000s, the change between the two decades and the smoothed distributions of changes (vertical lines show mean changes, with MAC indicated). Different stand-age classes are shown in each row as follows: **a-d**, 0-20 years; **e-h**, 20-40 years; **i-l**, 40-60 years; **m-p**, 60-80 years; **q-t**, 80-100 years.



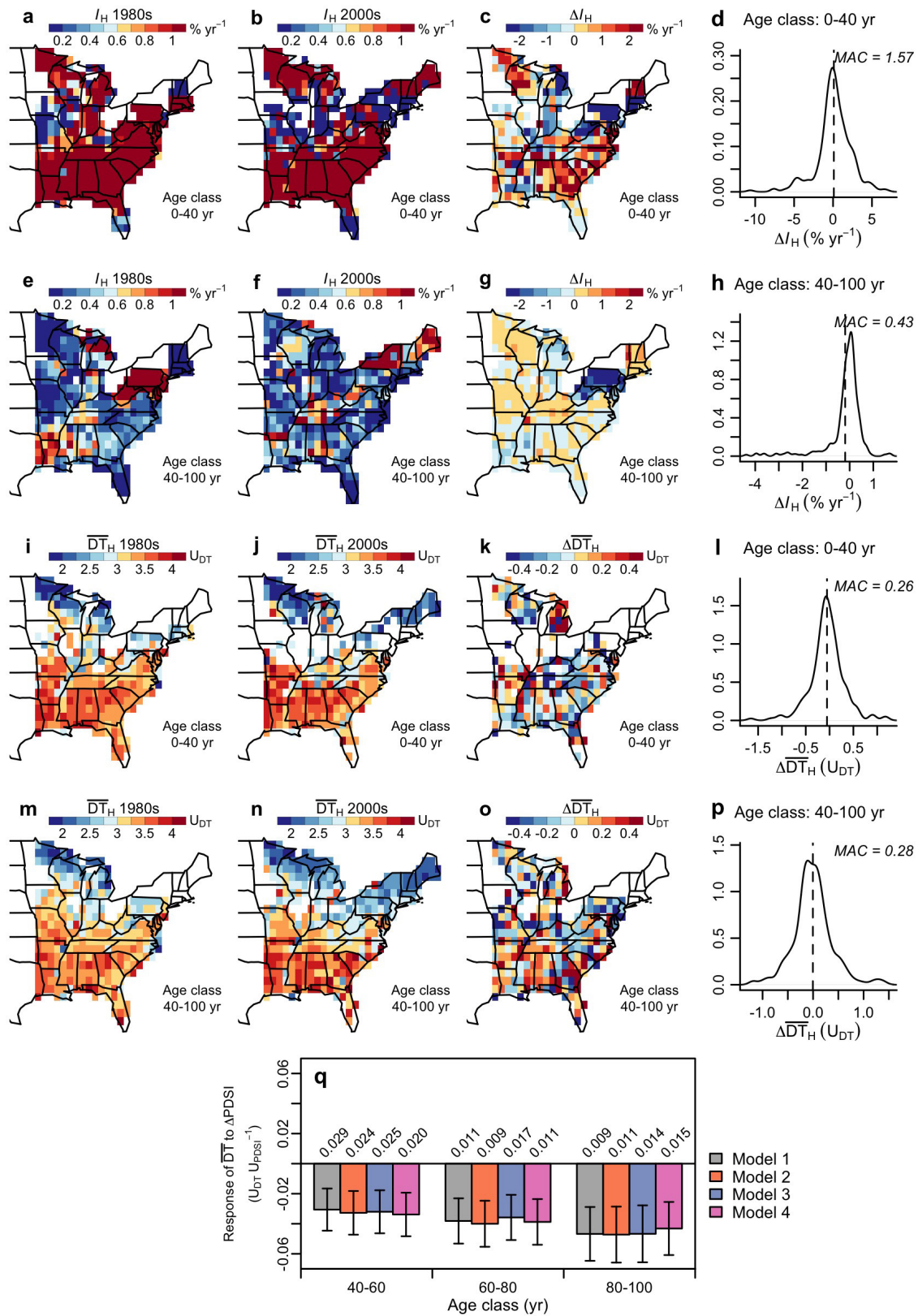
Extended Data Figure 4 | Sample sizes and slopes for grid-cell and stand-scale analyses. a, Summary of plot ages and sample sizes by age class for FIA data used in SAR models and other analyses (Figs 2, 3 and Extended Data Figs 7, 8). The bottom row of the table refers to the number of remeasured inventory plots used in stand-scale analyses (**d**, **e**), in which individual plots were tracked over time (Supplementary Methods 4). All other rows in the table refer to grid-cell-scale analyses that control for stand age by comparing plots in a given age class in the 1980s to plots in the same grid cell and age class in the 2000s (**b**, **c**). The number of grid cells varies across age classes because a grid cell was only included in the analysis for a given age class if the grid cell included at least five plot records that met our filtering criteria in both the 1980s and 2000s

(see Methods). **b**, SAR model slopes quantifying the mean grid-cell-scale response of $\Delta\overline{DT}$ (in DT units, U_{DT} , in which DT increases with drought tolerance and ranges from 1 to 5) to $\Delta PDSI$ (in PDSI units, U_{PDSI}) from the 1980s to 2000s. **c**, SAR model slopes quantifying the mean grid-cell-scale response of ΔAGB ($Mg\ ha^{-1}$) to $\Delta PDSI$ and $\Delta\overline{DT}$ from the 1980s to 2000s. **d**, **e**, Similar to **b**, **c**, but for stand-level SAR models fit to remeasured plots (see Supplementary Methods 4 for details). All SAR slopes are partial regression coefficients that control for changes in growing-season length, changes in community-mean shade tolerance, and spatial autocorrelation (see Methods). Error bars are s.e.m. of slopes, with P values shown outside of the bars: * $P \leq 0.05$; ** $P \leq 0.01$.



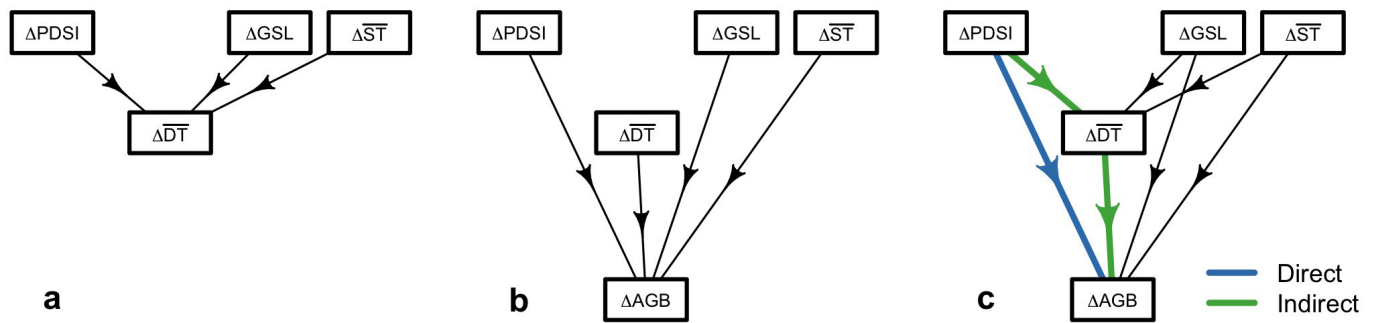
Extended Data Figure 5 | Community-mean shade tolerance in the 1980s and 2000s in the eastern USA. \overline{ST} was calculated within 1° grid cells and 20-year stand-age classes that contained at least five FIA inventory plots that satisfied filtering criteria (see Methods). From left, the columns show values of \overline{ST} for the 1980s, the 2000s, the change between the two decades and the smoothed distributions of changes (vertical lines

show mean changes, with MAC indicated). Different stand-age classes are shown in each row as follows: **a–d**, 0–20 years; **e–h**, 20–40 years; **i–l**, 40–60 years; **m–p**, 60–80 years; **q–t**, 80–100 years. \overline{ST} units (U_{ST}) are on the scale of the species shade tolerance (ST) index, which increases from 1 (very intolerant) to 5 (very tolerant); see Methods.

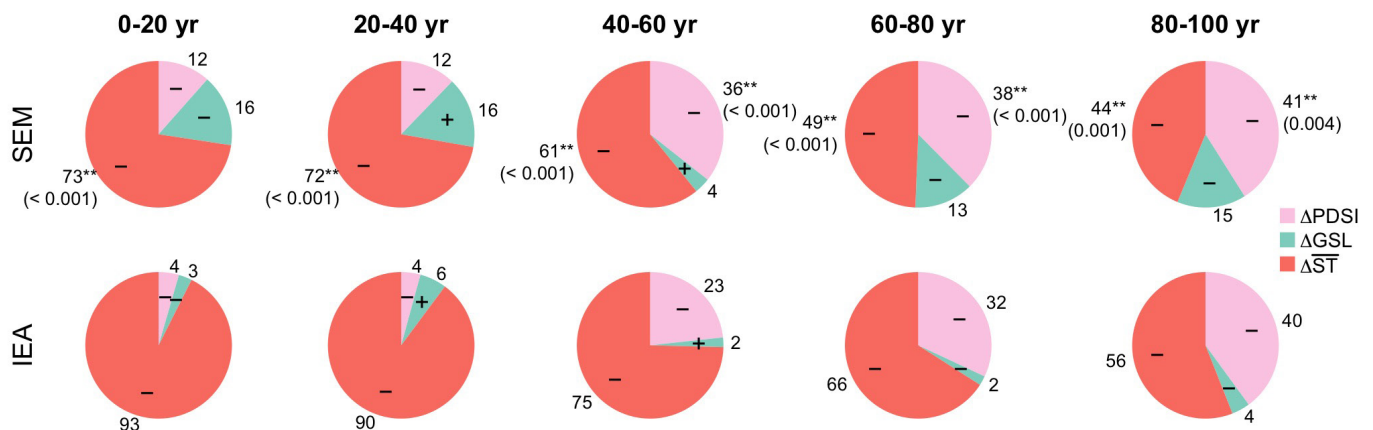


Extended Data Figure 6 | Harvest intensity and mean drought tolerance of harvested trees in the 1980s and 2000s in the eastern USA. Harvest intensity I_H (per cent AGB harvested per year, units $\% \text{ yr}^{-1}$) and \overline{DT}_H (units U_{DT}) were calculated within 1° grid cells and 20-year stand-age classes for analysis (Supplementary Methods 3), but are shown here in aggregated age classes (0–40 and 40–100 years) because patterns were similar among 20-year age classes within each aggregated class. **a–p**, From left, the columns show values of I_H (**a–h**) or \overline{DT}_H (**i–p**) for the 1980s, the 2000s, the change between the two decades, and the smoothed distributions of changes (vertical lines show mean changes, with MAC indicated). Different stand-age classes are shown in each

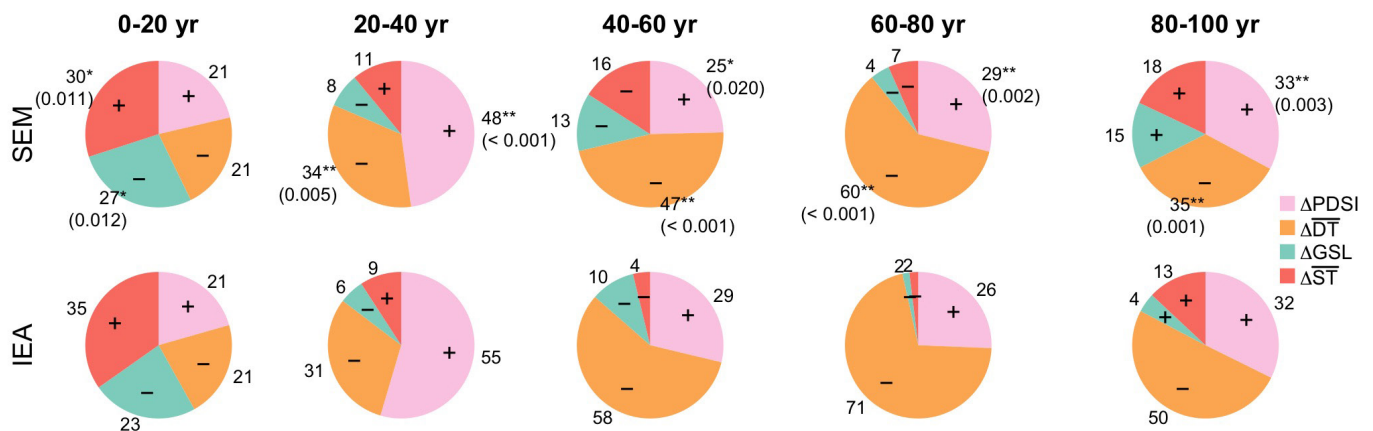
row as follows: **a–d**, **i–l**, 0–40 years; **e–h**, **m–p**, 40–100 years. **q**, Slopes of $\Delta \overline{DT}_H$ versus ΔPDSI from four versions of spatial regression models with different tree- and plot-filtering criteria and different approaches to modelling harvest effects (see Supplementary Methods 3 for details). Model 1 corresponds to Extended Data Fig. 4b (note that only the three significant age classes from Extended Data Fig. 4b are presented here: 40–60, 60–80 and 80–100 years). The similar slopes and significance levels (P values are shown above each bar) from the four models suggest that the estimated response of $\Delta \overline{DT}_H$ to ΔPDSI is robust to including or excluding effects of tree harvest. Error bars are s.e.m. of slopes.



d. $\Delta\overline{DT}$ model



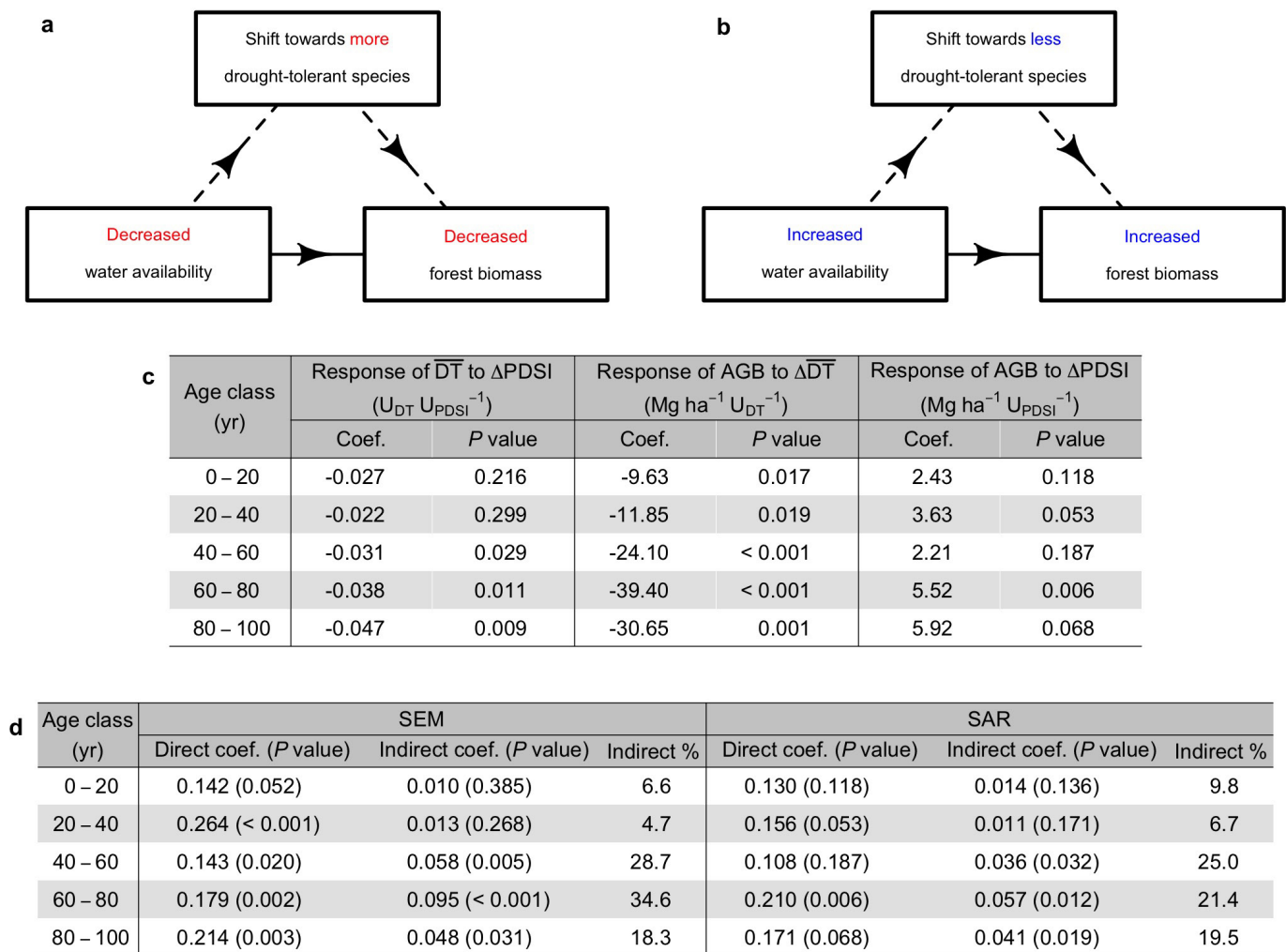
e. $\Delta\overline{AGB}$ model



Extended Data Figure 7 | Structural equation model and independent effects analysis results for the response variables $\Delta\overline{DT}$ and $\Delta\overline{AGB}$.

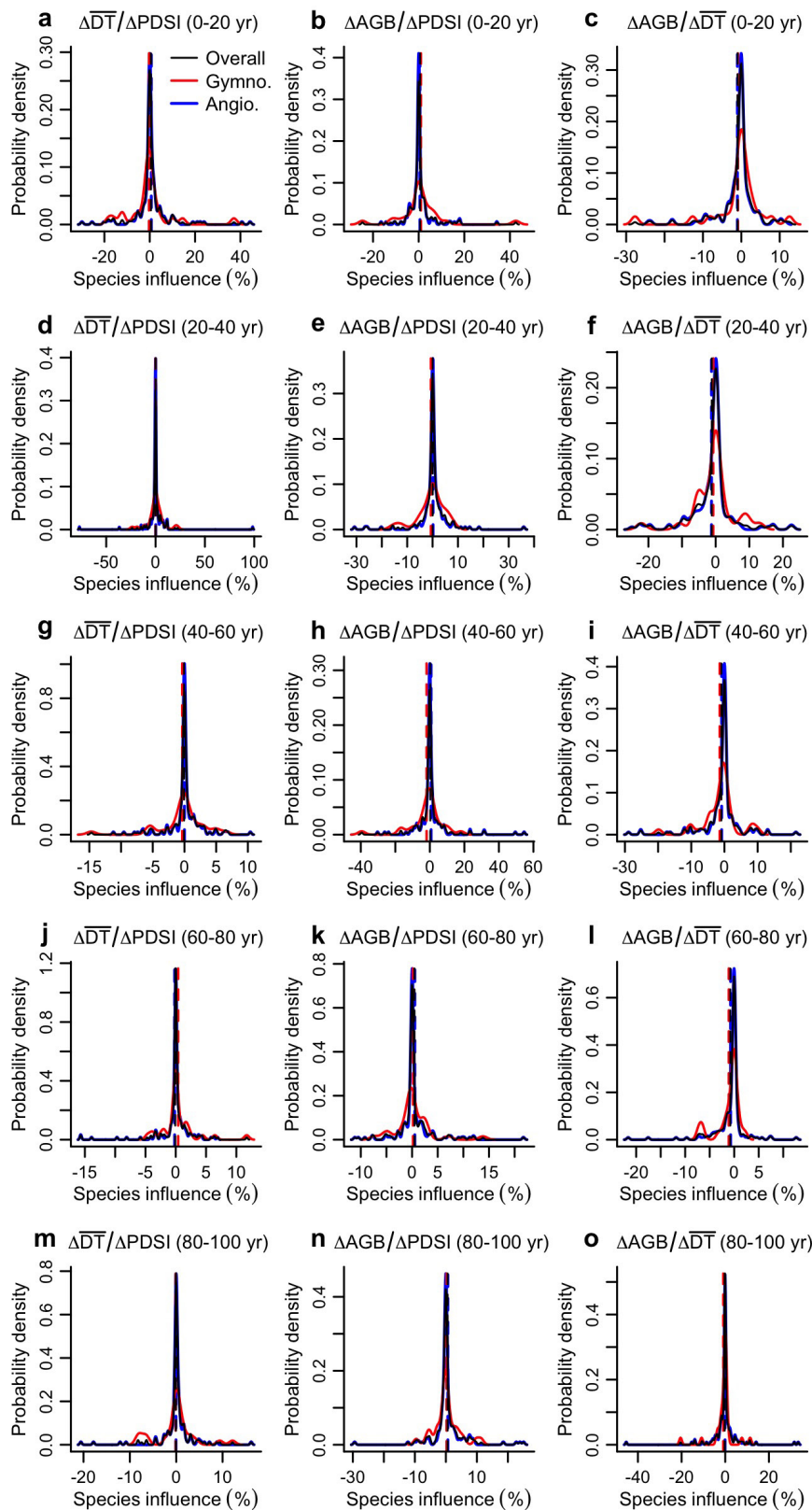
a, SEM structure for the $\Delta\overline{DT}$ model. **b**, SEM structure for the $\Delta\overline{AGB}$ model. **c**, SEM structure for the alternative $\Delta\overline{AGB}$ model that includes both direct and indirect effects of $\Delta\overline{PDSI}$ on $\Delta\overline{AGB}$ (see Methods and Extended Data Fig. 8a, b for further explanation of direct and indirect effects). **d**, Per cent contributions of $\Delta\overline{PDSI}$, $\Delta\overline{GSL}$ and $\Delta\overline{ST}$ to the explained variation in $\Delta\overline{DT}$. **e**, Per cent contributions of $\Delta\overline{PDSI}$, $\Delta\overline{GSL}$, $\Delta\overline{ST}$ and $\Delta\overline{DT}$ to the explained variation in $\Delta\overline{AGB}$ (SEM results are for

the model structure shown in **b**). Direct and indirect effects from **c** are reported in Extended Data Fig. 8d. Both SEM and IEA provide variance-partitioning estimates. In addition, SEM provides significance tests for explanatory variables. Significant P values ($P \leq 0.05$) are shown in the parentheses below the corresponding percentage contribution, * $P \leq 0.05$; ** $P \leq 0.01$. See Methods for details of SEM and IEA analyses. Positive and negative signs in **d**, **e** indicate the signs of the SEM and IEA coefficients, which are consistent with the signs of SAR model coefficients (Extended Data Figs 4 and 8).



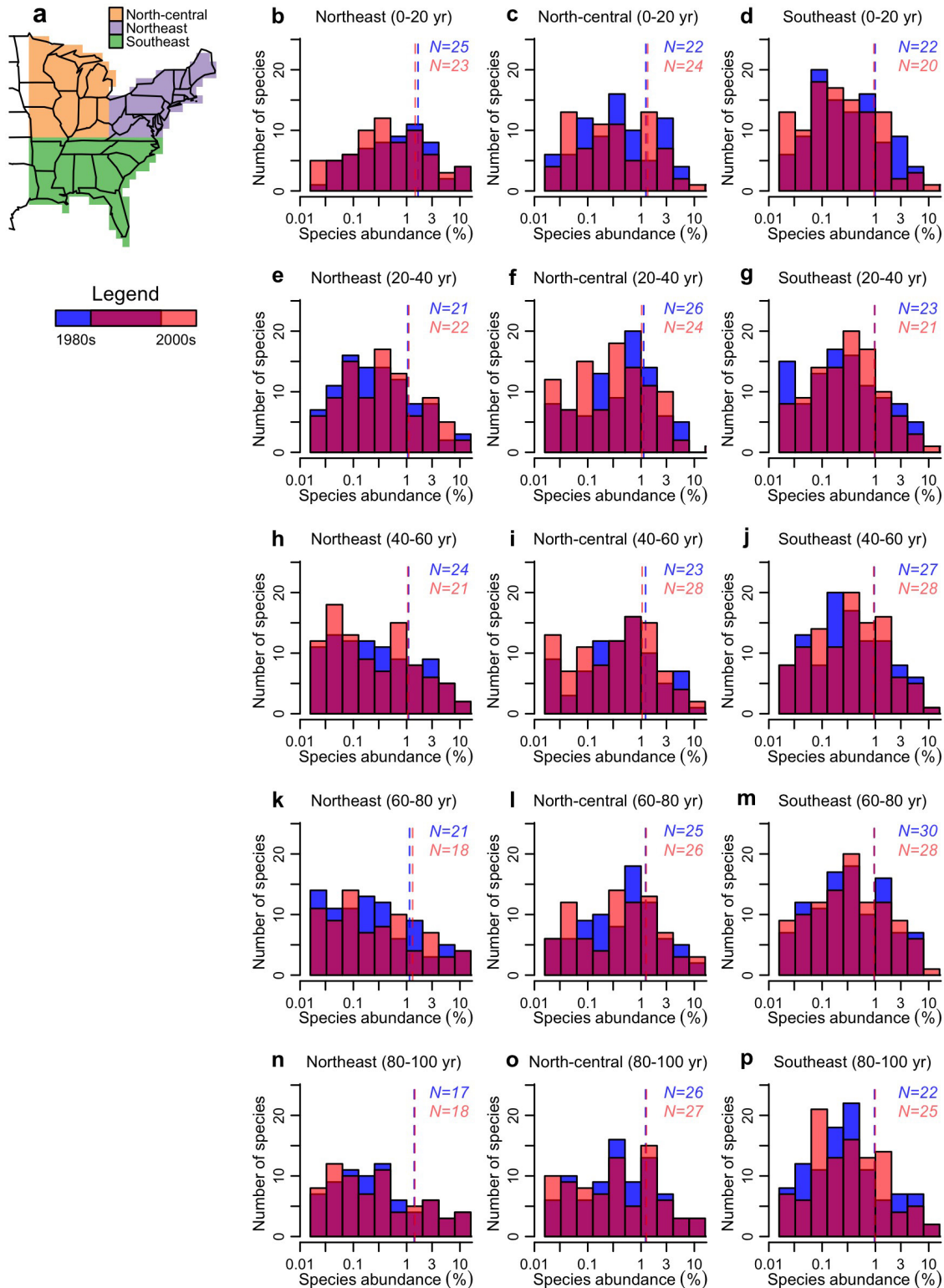
Extended Data Figure 8 | Conceptual model and supporting evidence for direct and indirect effects of changes in water availability on forest biomass. **a**, Conceptual model of direct (solid arrow) and indirect (dashed arrows) effects of decreasing water availability ($\Delta PDSI < 0$) on forest biomass. **b**, Conceptual model of direct (solid arrow) and indirect (dashed arrows) effects of increasing water availability ($\Delta PDSI > 0$) on forest biomass. The conceptual model in **a**, **b** is supported by our results (**c**, **d**), which show that the response of forest biomass to $\Delta PDSI$ is amplified by indirect effects; for example, if water availability decreases ($\Delta PDSI < 0$), then biomass decreases owing to direct effects (for example, physiological tree-level responses) as well as indirect effects (shifts in species composition towards more drought-tolerant but lower-biomass species). **c**, Slopes and *P* values from SAR models (equations (1) and (2)) for

different stand-age classes (1° grid-cell-scale results, as in Extended Data Fig. 4b, c). Slopes are partial regression coefficients, so the slope labelled 'response of AGB to $\Delta PDSI$ ' is the direct effect of $\Delta PDSI$ on ΔAGB (controlling for $\Delta \overline{DT}$ and other covariates), and the indirect effect is estimated by the product of the other two slopes (see Methods). **d**, Direct and indirect effects of $\Delta PDSI$ on AGB estimated from SEM (Extended Data Fig. 7c) and SAR models (see Methods). Sample sizes in these analyses (number of grid cells) are: 171, 247, 271, 271, and 193 for age classes from 0–20 to 80–100, respectively (as shown in Extended Data Fig. 4a). For consistency with SEM, SAR slopes are standardized in **d** (standard deviation units) but are otherwise equivalent to SAR coefficients in **c** and in Fig. 3. The percentages in **d** are calculated as $100 \times \text{indirect} / (\text{direct} + \text{indirect})$.



Extended Data Figure 9 | Distributions of species influence on estimated responses of community-mean drought tolerance (\overline{DT}) to $\Delta PDSI$, AGB to $\Delta PDSI$, and AGB to $\Delta \overline{DT}$. Species influence (x axis) is the per cent change in SAR model slopes owing to including an individual species in the analysis (see Supplementary Methods 6, 7 for details). Probability density (y axis) is the relative frequency of species with a given per cent influence. Black curves are for all species, and red and blue curves are for gymnosperm and angiosperm species, respectively. Vertical lines

are means. The distributions show that most species have little influence on the SAR slopes, some species have large influence, and gymnosperms and angiosperms have similar degrees of influence. From left, the columns show species influences on parameter d_P from equation (1), and species influences on parameters a_P and a_D from equation (2). Different stand-age classes are shown in each row as follows: a–c, 0–20 years; d–f, 20–40 years; g–i, 40–60 years; j–l, 60–80 years; m–o, 80–100 years.



Extended Data Figure 10 | Species abundance distributions in the northeastern, north-central and southeastern USA in the 1980s and 2000s. **a**, Map of the three mentioned sub-regions of the USA. **b–p**, Species abundance distributions (number of species in different abundance intervals) for each age class (rows) in each sub-region (columns). Blue and red bars represent numbers of species in the 1980s and 2000s, respectively. Purple indicates overlap of the two bars. Abundance is quantified as the percentage of total above-ground biomass comprised by

a given species in a given sub-region and age class. Abundance intervals are on a logarithmic (base-2) scale with the following lower limits: 2^{-6} , 2^{-5} , 2^{-4} , ..., 2^3 (that is, 0.015625%, 0.03125%, 0.0625%, 0.125%, 0.25%, 0.5%, 1%, 2%, 4% and 8%). Species with abundances of less than 2^{-6} (0.015625%) were excluded. Vertical lines are mean abundances of species with abundance $\geq 0.015625\%$. The number of common species (defined as species with an abundance $\geq 1\%$) is given for the 1980s (blue text) and 2000s (red text).

A new class of synthetic retinoid antibiotics effective against bacterial persisters

Wooseong Kim¹, Wenpeng Zhu², Gabriel Lambert Hendricks¹, Daria Van Tyne^{3,4}, Andrew D. Steele^{5,6}, Colleen E. Keohane^{5,6}, Nico Fricke², Annie L. Conery^{7,8}, Steven Shen¹, Wen Pan¹, Kiho Lee¹, Rajmohan Rajamuthiah¹, Beth Burgwyn Fuchs¹, Petia M. Vlahovska⁹, William M. Wuest^{5,6}, Michael S. Gilmore^{3,4}, Huajian Gao², Frederick M. Ausubel^{7,8} & Eleftherios Mylonakis¹

A challenge in the treatment of *Staphylococcus aureus* infections is the high prevalence of methicillin-resistant *S. aureus* (MRSA) strains and the formation of non-growing, dormant ‘persister’ subpopulations that exhibit high levels of tolerance to antibiotics^{1–3} and have a role in chronic or recurrent infections^{4,5}. As conventional antibiotics are not effective in the treatment of infections caused by such bacteria, novel antibacterial therapeutics are urgently required. Here we used a *Caenorhabditis elegans*–MRSA infection screen⁶ to identify two synthetic retinoids, CD437 and CD1530, which kill both growing and persister MRSA cells by disrupting lipid bilayers. CD437 and CD1530 exhibit high killing rates, synergism with gentamicin, and a low probability of resistance selection. All-atom molecular dynamics simulations demonstrated that the ability of retinoids to penetrate and embed in lipid bilayers correlates with their bactericidal ability. An analogue of CD437 was found to retain anti-persister activity and show an improved cytotoxicity profile. Both CD437 and this analogue, alone or in combination with gentamicin, exhibit considerable efficacy in a mouse model of chronic MRSA infection. With further development and optimization, synthetic retinoids have the potential to become a new class of antimicrobials for the treatment of Gram-positive bacterial infections that are currently difficult to cure.

We used an established automated high-throughput *C. elegans*–MRSA killing assay in 384-well plates⁶ to screen approximately 82,000 small synthetic molecules, and identified 185 compounds that significantly decreased the ability of MRSA to kill the nematodes (Fig. 1a, Supplementary Table 1, Supplementary Methods). Two of these 185 compounds, the synthetic retinoids CD437 and CD1530 (vitamin A analogues; Fig. 1b), were selected for further investigation because they have similar structures and have been studied previously for their therapeutic potential^{7–10}.

CD437 and CD1530 exhibit potent *in vitro* bactericidal activity against MRSA strain MW2; after two hours, levels of MW2 were below the limit of detection (minimum inhibitory concentration (MIC) 1 µg ml^{−1}; Fig. 1c, Extended Data Fig. 1a, Supplementary Table 2). *In vivo*, CD437 or CD1530 at concentrations above their *in vitro* MICs protected 100% of *C. elegans* against MW2-induced death (Fig. 1d). CD437 and CD1530 also exhibited potent activity against a panel of clinical *S. aureus* and *Enterococcus faecium* strains, but not against Gram-negative species (Supplementary Table 2). In addition, adapalene, a structural analogue of CD437 and CD1530 and a potential ovarian cancer drug⁸ (Fig. 1b), also exhibited significant anti-staphylococcal activity (MIC 2 µg ml^{−1}) and prevented the MRSA-induced death of *C. elegans*. However, adapalene, another analogue and a US Food and Drug Administration (FDA)-approved acne

therapeutic¹¹, was ineffective against MRSA (Fig. 1a–d, Extended Data Fig. 1a, Supplementary Table 2).

We were unable to obtain retinoid-resistant mutants by plating 10¹⁰ colony-forming units (CFU) of *S. aureus* MW2 on agar containing 2.5×, 5× or 10× MIC of CD437, CD1530 or adapalene. Similarly, serial passage of two independent *S. aureus* MW2 cultures (SP1 and SP2) for 100 days in sub-MIC levels of CD437 yielded only putative mutants with twofold greater resistance to CD437, CD1530 or adapalene, whereas serial passage in ciprofloxacin for 100 days (Fig. 1e) or daptomycin for 15 days (Extended Data Fig. 1b) generated strains that were 256-fold and tenfold more resistant, respectively. The MW2 cultures that exhibited modest retinoid resistance contained mutations in the genes *gras*, *yjbH* and *manA* (Fig. 1f, Supplementary Tables 3–5, Supplementary Discussion), which encode products related to membrane physiology^{12–16}. Consistent with this finding, CD437, CD1530 and adapalene—but not adapalene—induced membrane permeabilization in MW2 (monitored by SYTOX Green uptake; Fig. 2a), and CD437 and CD1530 caused the formation of mesosome-like structures (observed by transmission electron microscopy; Fig. 2b), similar to those observed in *S. aureus* cells after treatment with antimicrobial peptides¹⁷. Moreover, CD437, CD1530 and adapalene disrupted the integrity of biomembrane-mimicking giant unilamellar vesicles (Fig. 2c, Supplementary Videos 1–5). These vesicles consist of a DOPC:DOPG lipid bilayer at a ratio of 7:3 (DOPC/G, 1,2-dioleoyl-*sn*-glycero-3-phosphocholine/glycerol), which mimics anionic bacterial membranes, and have been used to elucidate the mechanisms of action of daptomycin in *S. aureus*^{18,19}. Notably, however, CD437 and CD1530 did not lyse bacterial cells directly (Extended Data Fig. 1c).

To elucidate the molecular interactions between retinoids and the membrane lipid bilayers of *S. aureus*, we conducted all-atom molecular dynamics simulations using a lipid bilayer composed of 108 phosphatidylglycerol lipids, 72 lysyl-phosphatidylglycerol (Lys-PG) lipids and 10 diphosphatidylglycerol (DPG, also known as cardiolipin) lipids, which mimics the phospholipid composition of *S. aureus* membranes²⁰. These simulations showed that the carboxylic acid and the phenolic groups of CD437, CD1530 and adapalene anchor these retinoids to the surface of the membrane bilayer by binding persistently to hydrophilic lipid heads. As a result, the retinoids penetrate the bilayers and become embedded orthogonally to the lipid molecules in the outer membrane leaflet, inducing substantial perturbations of the membrane (Fig. 2d, e, Supplementary Videos 6–9). Similar results were obtained for molecular dynamics simulations of DOPC:DOPG (7:3) lipid bilayers used in the giant unilamellar vesicle experiments in Fig. 2c (Extended Data Fig. 2a, b, Supplementary Videos 10–13). In contrast to CD437, CD1530 and adapalene, adapalene does not penetrate the membrane owing to a high energy barrier (11.22 k_BT) and an unfavourable

¹Division of Infectious Diseases, Rhode Island Hospital, Warren Alpert Medical School of Brown University, Providence, Rhode Island 02903, USA. ²School of Engineering, Brown University, Providence, Rhode Island 02903, USA. ³Department of Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts 02114, USA. ⁴Department of Microbiology and Immunobiology, Harvard Medical School, Massachusetts 02115, USA. ⁵Department of Chemistry, Emory University, Atlanta, Georgia 30322, USA. ⁶Emory Antibiotic Resistance Center, Emory University, Atlanta, Georgia 30322, USA. ⁷Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁸Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁹Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, Illinois 60208, USA.

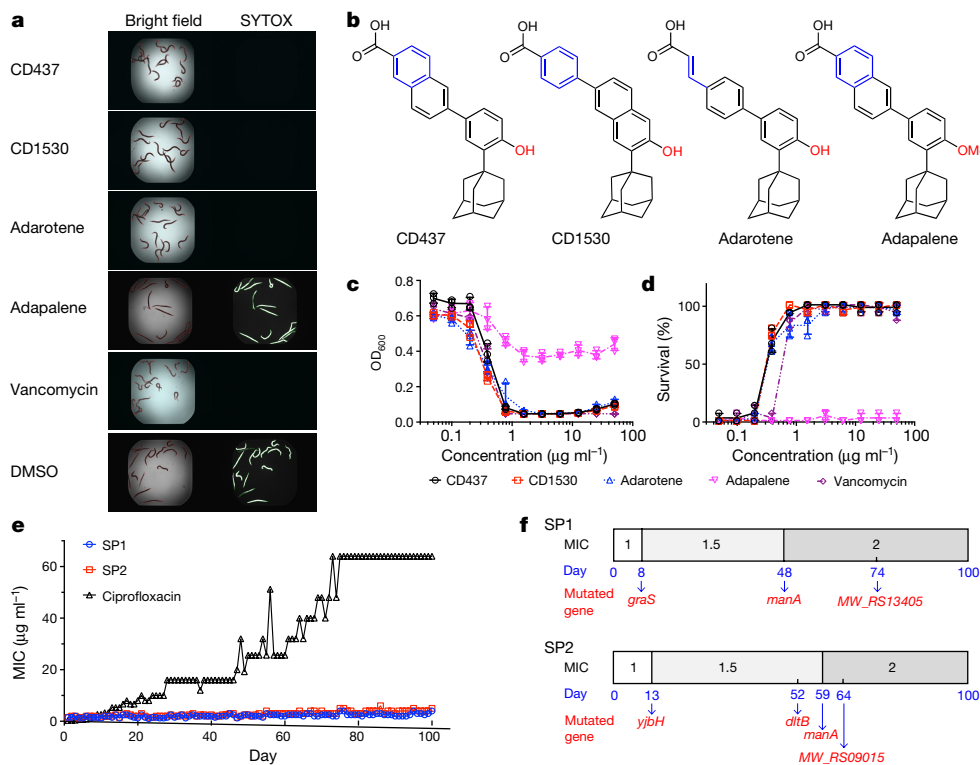


Figure 1 | Synthetic retinoids protect *C. elegans* from MRSA infection and inhibit MRSA growth without detectable mutant development. **a**, Images of MRSA-MW2-infected *C. elegans* in the presence of $10\mu\text{g ml}^{-1}$ retinoids, $10\mu\text{g ml}^{-1}$ vancomycin, or 1% DMSO as a control (see Supplementary Methods). Only dead worms stain with SYTOX Orange. Experiments were independently repeated three times with similar results. **b**, Chemical structures of synthetic retinoids. **c**, Growth of MW2 exposed to the five indicated compounds at various concentrations after 18 hours in tryptic soy broth. OD₆₀₀, optical density at 600 nm. **d**, Survival of *C. elegans* infected with MW2 in the presence of retinoids, normalized to *C.*

transfer energy ($3.16 k_B T$) (Fig. 2e, Supplementary Table 6), as the hydrophobic methoxy group does not bind to lipid heads (Fig. 2d, Supplementary Video 9). CD437-like retinoids can be metabolized in the liver by glucuronidation at carboxylic or hydroxyl groups²¹. Molecular dynamics simulations showed that two CD437 glucuronide metabolites also penetrate into lipid bilayers, exhibiting a similar penetration mechanism to that of CD437 (Extended Data Fig. 2a, c, Supplementary Videos 14, 15, Supplementary Discussion). In summary, these molecular dynamics simulations showed that two polar branch groups—a phenol and a carboxylate—have essential roles in membrane attachment and penetration, and that the membrane activity of retinoids (Fig. 2) directly correlates with their antibiotic activity (Fig. 1c, Extended Data Fig. 1a).

CD437 or CD1530—but not adarotene—also induced rapid permeabilization of MRSA-persister membranes (Extended Data Fig. 3a), and killed MRSA-persister cells (Fig. 3a, Extended Data Fig. 3b). They also completely eradicated persisters formed by 13 clinical isolates, including the multi-drug resistant strain VRS1 within 1 to 4 hours at 8–10× MIC (Fig. 3b, Extended Data Fig. 3c, d). Moreover, CD437 or CD1530 killed around 90% and 100% of persisters formed in MRSA biofilms at 16× MIC and 32× MIC, respectively (Extended Data Fig. 4). Compared with adarotene, CD437 and CD1530 can penetrate the membrane more efficiently owing to lower energy barriers and more favourable transfer energies (Fig. 2e); this is consistent with the observation that adarotene is inactive against persister cells. The results suggest that the planar aryl moiety of CD437 and CD1530 (highlighted in blue, Fig. 1b) rigidifies the carboxylic acid, which facilitates penetration into lipid bilayers, whereas the flexible cinnamoyl moiety of adarotene

elegans treated with DMSO. **c**, **d**, Individual data points ($n=3$ biologically independent experiments) and mean \pm s.d. are shown. **e**, Appearance of spontaneous CD437- and ciprofloxacin-resistant MW2 mutants over 100 days of serial passage in duplicate (SP1 and SP2) (see Supplementary Methods). **f**, Appearance of mutations on specific days in the indicated genes in SP1 and SP2 in **e** (see Supplementary Methods). The modest increase in the MIC of CD437 against MRSA during serial passage was confirmed by remeasuring MICs using three colonies from aliquots of each passage that had been stored at -80°C . Mutated genes are indicated on the day at which the mutations were first detected.

fails to orient the carboxylate appropriately, thereby decreasing membrane penetration (Fig. 2e).

CD437 or CD1530 exhibited significant synergism with gentamicin against both MRSA growing and persister cells (Fig. 3c, Supplementary Table 7, Extended Data Fig. 4). This is most probably a consequence of the increased passive diffusion of gentamicin through the bacterial cell membranes that have been physically damaged by the retinoids, which is mechanistically distinct from the observed synergism between gentamicin and ionophores²² (Extended Data Fig. 5, Supplementary Discussion).

Although membrane-targeting agents often cause toxicity in mammals²³, CD437, CD1530 and adarotene are relatively non-toxic, exhibiting median haemolytic concentrations (HC₅₀) of greater than $32\mu\text{g ml}^{-1}$ (Extended Data Fig. 6a). CD437, CD1530 and adarotene were more toxic to human hepatoma HepG2 cells (median lethal concentration (LC₅₀) 3–5 $\mu\text{g ml}^{-1}$) than to normal human primary hepatocytes (LC₅₀ $\geq 20\mu\text{g ml}^{-1}$), or to primary renal proximal tubule epithelial cells or adult normal human epidermal keratinocytes at $8\mu\text{g ml}^{-1}$ (Extended Data Fig. 6b), a concentration at which CD437 and CD1530 completely eradicated MRSA persisters (Fig. 3a). These data are consistent with previous results showing that CD437 exhibits selective toxicity towards cancer cells¹⁰. None of the three retinoids inhibited the human ether-a-go-go related (hERG) potassium channels that are critical for cardiac action potential repolarization at $25\mu\text{M}$ (Extended Data Fig. 6c) and did not show significant genotoxic potential (Supplementary Table 8).

To evaluate the effects of the CD437-like retinoid branch groups on antimicrobial activity and the possibility of further structural

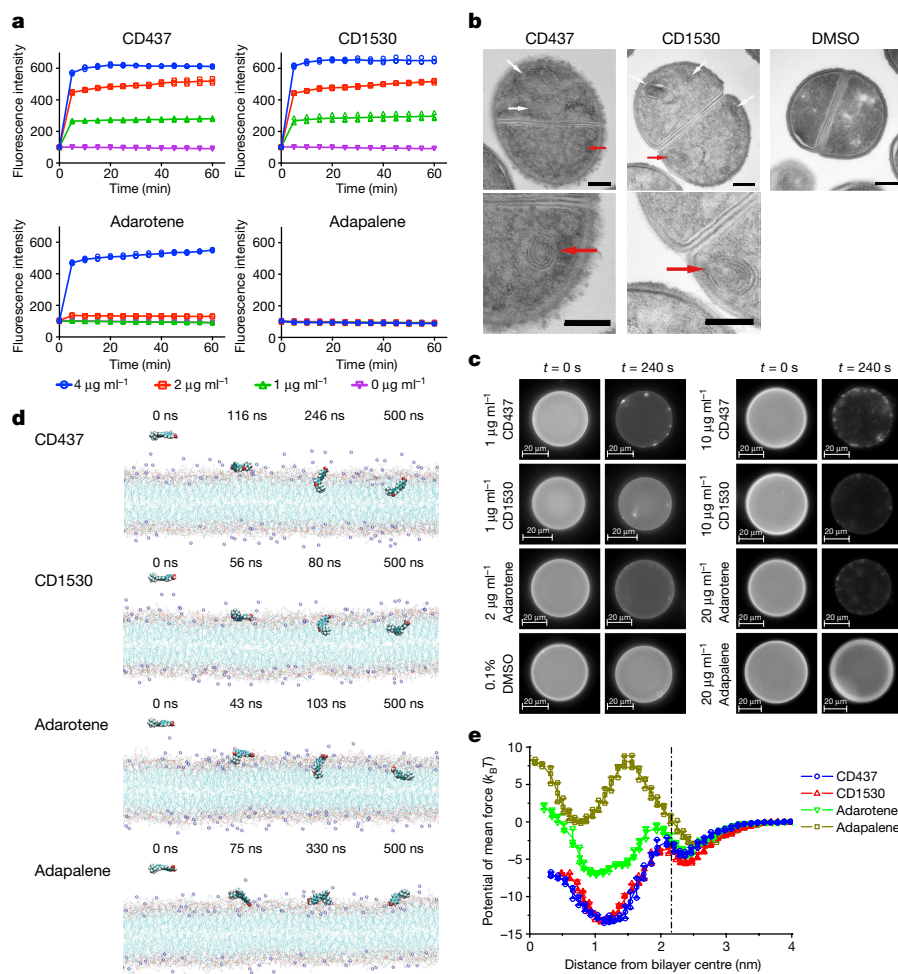


Figure 2 | CD437, CD1530 and adapalene disrupt membrane lipid bilayers. **a**, Uptake of SYTOX Green ($\lambda_{\text{ex}} = 485 \text{ nm}$, $\lambda_{\text{em}} = 525 \text{ nm}$) by exponential-phase *S. aureus* MW2 cells treated with retinoids. Individual data points ($n = 3$ biologically independent samples) and means are shown. Error bars not shown for clarity. **b**, Transmission electron micrographs showing mesosome-like structures (white and red arrows; enlarged in bottom images) in $10\times$ MIC retinoid-treated cells and DMSO control. Scale bars, 200 nm . **c**, Changes in giant unilamellar vesicles (DOPC:DOPG, 7:3) labelled with 18:1 Liss Rhod PE (0.05%) treated with retinoids or with 0.1% DMSO, monitored using fluorescence microscopy ($40\times$ objective, $\lambda_{\text{ex}} = 460 \text{ nm}$, $\lambda_{\text{em}} = 483 \text{ nm}$). Liss Rhod PE, 1,2-dioleoyl-*sn*-glycero-3-phosphoethanolamine-*N*-(lissamine rhodamine B sulfonyl) (ammonium salt). In **b**, **c**, experiments were independently repeated

optimization with respect to antimicrobial and toxicity profiles, we synthesized 16 analogues of CD437 (Extended Data Fig. 7a). Subsequent analysis of their structure–activity relationships supported the putative mode of action by which the synthetic retinoids disrupt Gram-positive bacterial membranes, and demonstrated that the antimicrobial activity and cytotoxicity of synthetic retinoids can be modulated by the polarity of the branch groups (Extended Data Figs 7–9, Supplementary Videos 16, 17, Supplementary Discussion). In particular, analogue 2, which has a less polar primary alcohol instead of the carboxylic acid group, retained bacterial activity against MRSA persisters (Fig. 4a, b), but showed significantly less haemolytic activity ($\text{HC}_{50} > 32 \mu\text{g ml}^{-1}$, Extended Data Fig. 8a) and less cytotoxicity in a panel of human cell lines ($\text{LC}_{50} \geq 31 \mu\text{g ml}^{-1}$) than did CD437 (Fig. 4c, Extended Data Fig. 6b). Analogue 2 also showed significantly reduced activity towards human hepatoma HepG2 cells, with LC_{50} values of $> 32 \mu\text{g ml}^{-1}$ (Fig. 4c). In addition, molecular dynamics simulations revealed that analogue 2 penetrates membrane lipid bilayers with similar energy profiles to those of CD437 (Extended Data Fig. 8d, e, Supplementary

Video 16), further establishing that the extent of membrane penetration inferred from molecular dynamics simulations correlates with antimicrobial activity. In summary, the structure–activity relationships verified that persistent attachment to lipid heads by the two polar branch groups is critical for antimicrobial activity, and that antimicrobial activity and lack of cytotoxicity can be optimized by simple modifications to the polar branch groups. Notably, analogue 2 also exhibited favourable pharmacokinetic profiles after intraperitoneal administration of a single dose of 20 mg kg^{-1} , with a maximum plasma concentration of around $10 \mu\text{g ml}^{-1}$ and an elimination half-life of 4.5 hours (Extended Data Fig. 8f). By contrast, adapalene is excreted rapidly^{21,24}. Analogue 2 showed no detectable hepatic or renal toxicity in mice at intraperitoneal doses of up to 80 mg kg^{-1} (the highest tested dose) every 12 hours for 3 days (Extended Data Fig. 8g).

Finally, we evaluated the efficacy of both analogue 2 and the combination of analogue 2 and gentamicin in a mouse deep-seated thigh MRSA infection model, which mimics human deep-seated chronic infections². Consistent with previous findings², a combination of vancomycin and

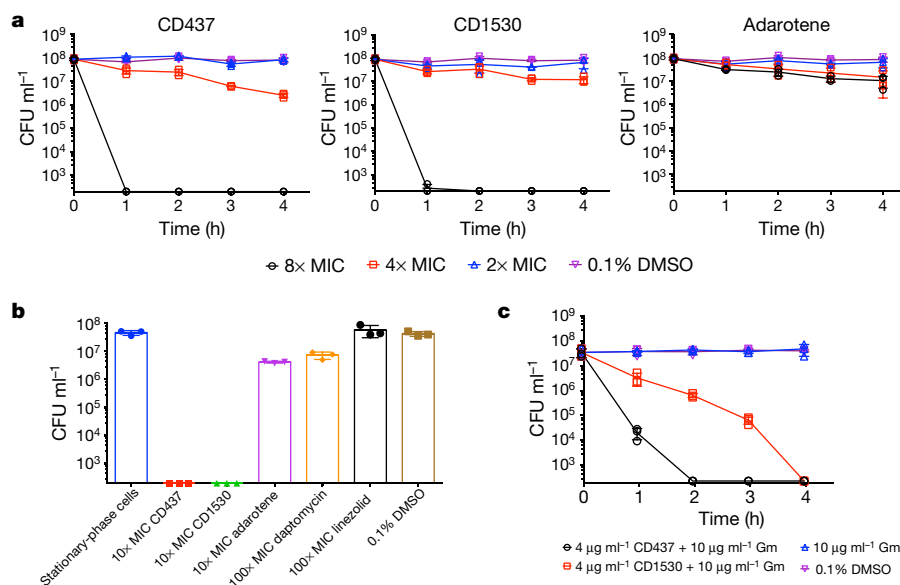


Figure 3 | CD437 or CD1530 alone or in combination with gentamicin are effective against persisters. **a**, **b**, Viability of stationary-phase *S. aureus* MW2 (**a**) or *S. aureus* VRS1 (**b**) when treated with the indicated concentrations of each retinoid for 4 hours. **c**, Viability upon treatment of

S. aureus MW2 persisters with the indicated concentrations of retinoids in combination with gentamicin (Gm). In **a**–**c**, the data points on the x axis are below the level of detection (2×10^2 CFU ml⁻¹). Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown.

gentamicin did not significantly reduce MRSA abundance (Fig. 4d) even though MW2 is sensitive to both antibiotics, which suggests that the bacterial cells in this infection model are persisters. As shown in Fig. 4d, 80 mg kg⁻¹ of analogue 2 alone led to an approximately fourfold

decrease ($P < 0.001$) in MRSA abundance, and 40 or 80 mg kg⁻¹ of analogue 2 in combination with 30 mg kg⁻¹ gentamicin resulted in approximately 14-fold ($P < 0.001$) and approximately 23-fold decreases ($P < 0.001$) in bacterial burden, respectively. Similarly, CD437 alone or

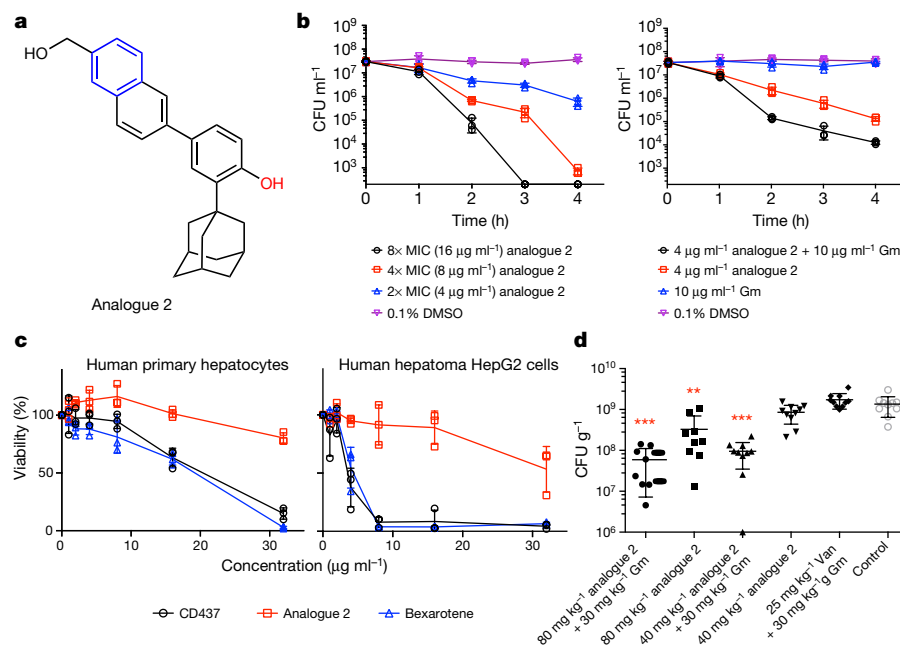


Figure 4 | Analogue 2 retains antimicrobial activity against MRSA persisters and has improved cytotoxicity compared with CD437. **a**, Chemical structure of analogue 2. **b**, Viability of *S. aureus* MW2 persisters treated with analogue 2 alone or in combination with gentamicin (Gm). Data points on the x axis were below the level of detection (2×10^2 CFU ml⁻¹). **c**, Viability of normal human primary hepatocytes and human hepatoma (HepG2) cells treated with retinoids in serum-free medium for 24 hours, based on the absorbance readings at 450 nm taken 4 hours after adding the tetrazolium dye WST-1. The FDA-approved antineoplastic retinoid bexarotene was used as a control. **b**, **c**, Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown. **d**, Efficacy of analogue 2 alone or in combination with gentamicin

in a deep-seated mouse thigh infection model. Each group of MW2-infected neutropenic mice ($n = 10$ biologically independent animals) was treated with the indicated doses of analogue 2 intraperitoneally (i.p.) alone or in combination with 30 mg kg⁻¹ subcutaneous (s.c.) gentamicin (Gm), a combination of 25 mg kg⁻¹ vancomycin (Van, i.p.) and 30 mg kg⁻¹ gentamicin (s.c.) or control (5% Kolliphor + 5% ethanol, i.p.) every 12 hours for 3 days beginning 24 hours after infection. At 12 hours after the last treatment, mice were euthanized and their thighs were excised and homogenized. CFUs from each mouse thigh are plotted as individual points. The mean \pm s.d. is shown. Statistical differences between control and antibiotic treatment groups were analysed by one-way ANOVA and post hoc Tukey test (** $P = 0.0002$, *** $P < 0.0001$).

in combination with gentamicin also exhibited efficacy in the MRSA mouse deep-seated thigh infection model (Extended Data Fig. 10). These results suggest that a combination of analogue 2 and gentamicin or CD437 and gentamicin might be an effective strategy to enhance the efficacy and reduce the toxicity of aminoglycosides²⁵ in the treatment of chronic Gram-positive infections.

Despite the potential advantages of membrane-active antimicrobials such as the retinoids described here—including fast killing, low probability of developing resistance, and anti-persister activity—the major obstacle for developing retinoids as therapeutics is their potential cytotoxicity, which is a matter of considerable debate^{23,26}. Nevertheless, we have identified a specific chemotype of membrane-active synthetic retinoids that are relatively selective for bacterial membranes and exhibit a high level of activity towards MRSA persister cells; these findings are notable because the development of appropriate antibiotics for persisters is an important unmet need. Although a limited analysis of structure–activity relationships showed that modification of the retinoid branch groups can result in improved cytotoxicity profiles while retaining anti-persister activity, it is important to acknowledge that the long term-potential of further chemical optimization of retinoids to develop non-toxic antimicrobials is currently unknown. However, considering the fact that the bioactivity of retinoids can be improved by modifying both the backbone and branch groups, and that approximately 4,000 retinoid analogues have been synthesized so far²⁶, our results warrant further development of synthetic retinoids as potential therapeutics for hard-to-treat infectious diseases caused by antibiotic-resistant or persistent Gram-positive pathogens.

Data Availability All data are available within the paper and its Supplementary Information.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 August 2016; accepted 26 February 2018.

Published online 28 March 2018.

- Allison, K. R., Brynildsen, M. P. & Collins, J. J. Metabolite-enabled eradication of bacterial persisters by aminoglycosides. *Nature* **473**, 216–220 (2011).
- Conlon, B. P. *et al.* Activated ClpP kills persisters and eradicates a chronic biofilm infection. *Nature* **503**, 365–370 (2013).
- Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
- Lew, D. P. & Waldvogel, F. A. Osteomyelitis. *Lancet* **364**, 369–379 (2004).
- Baddour, L. M. *et al.* Infective endocarditis in adults: diagnosis, antimicrobial therapy, and management of complications: a scientific statement for healthcare professionals from the American Heart Association. *Circulation* **132**, 1435–1486 (2015).
- Rajamuthiah, R. *et al.* Whole animal automated platform for drug discovery against multi-drug resistant *Staphylococcus aureus*. *PLoS ONE* **9**, e89189 (2014).
- Altucci, L., Leibowitz, M. D., Ogilvie, K. M., de Lera, A. R. & Gronemeyer, H. RAR and RXR modulation in cancer and metabolic disease. *Nat. Rev. Drug Discov.* **6**, 793–810 (2007).
- Valli, C. *et al.* Atypical retinoids ST1926 and CD437 are S-phase-specific agents causing DNA double-strand breaks: significance for the cytotoxic and antiproliferative activity. *Mol. Cancer Ther.* **7**, 2941–2954 (2008).
- Tang, X.-H. *et al.* Combination of bexarotene and the retinoid CD1530 reduces murine oral-cavity carcinogenesis induced by the carcinogen 4-nitroquinoline 1-oxide. *Proc. Natl Acad. Sci. USA* **111**, 8907–8912 (2014).
- Han, T. *et al.* The antitumor toxin CD437 is a direct inhibitor of DNA polymerase α . *Nat. Chem. Biol.* **12**, 511–515 (2016).
- Irby, C. E., Yentzer, B. A. & Feldman, S. R. A review of adapalene in the treatment of acne vulgaris. *J. Adolesc. Health* **43**, 421–424 (2008).
- Meehl, M., Herbert, S., Götz, F. & Cheung, A. Interaction of the GraRS two-component system with the VraFG ABC transporter to support vancomycin-intermediate resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **51**, 2679–2689 (2007).
- Yang, S.-J. *et al.* The *Staphylococcus aureus* two-component regulatory system, GraRS, senses and confers resistance to selected cationic antimicrobial peptides. *Infect. Immun.* **80**, 74–81 (2012).
- Elbaz, M. & Ben-Yehuda, S. The metabolic enzyme ManA reveals a link between cell wall integrity and chromosome morphology. *PLoS Genet.* **6**, e1001119 (2010).
- Falord, M., Mäder, U., Hiron, A., Débarbouillé, M. & Msadek, T. Investigation of the *Staphylococcus aureus* GraSR regulon reveals novel links to virulence, stress response and cell wall signal transduction pathways. *PLoS ONE* **6**, e21323 (2011).
- Göhring, N. *et al.* New role of the disulfide stress effector YjbH in β -lactam susceptibility of *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **55**, 5452–5458 (2011).
- Friedrich, C. L., Moyses, D., Beveridge, T. J. & Hancock, R. E. Antibacterial action of structurally diverse cationic peptides on gram-positive bacteria. *Antimicrob. Agents Chemother.* **44**, 2086–2092 (2000).
- Chen, Y.-F., Sun, T.-L., Sun, Y. & Huang, H. W. Interaction of daptomycin with lipid bilayers: a lipid extracting effect. *Biochemistry* **53**, 5384–5392 (2014).
- Ganewatta, M. S. *et al.* Bio-inspired resin acid-derived materials as antibacterial resistance agents with unexpected activities. *Chem. Sci.* **5**, 2011–2016 (2014).
- Piggot, T. J., Holdbrook, D. A. & Khalid, S. Electroporation of the *E. coli* and *S. aureus* membranes: molecular dynamics simulations of complex bacterial membranes. *J. Phys. Chem. B* **115**, 13381–13388 (2011).
- Sala, F. *et al.* Development and validation of a liquid chromatography–tandem mass spectrometry method for the determination of ST1926, a novel oral antitumor agent, adamantyl retinoid derivative, in plasma of patients in a Phase I study. *J. Chromatogr. B* **877**, 3118–3126 (2009).
- Farha, M. A., Verschoor, C. P., Bowdish, D. & Brown, E. D. Collapsing the proton motive force to identify synergistic combinations against *Staphylococcus aureus*. *Chem. Biol.* **20**, 1168–1178 (2013).
- Hurdle, J. G., O'Neill, A. J., Chopra, I. & Lee, R. E. Targeting bacterial membrane function: an underexploited mechanism for treating persistent infections. *Nat. Rev. Microbiol.* **9**, 62–75 (2011).
- Basma, H. *et al.* The synthetic retinoid ST1926 as a novel therapeutic agent in rhabdomyosarcoma. *Int. J. Cancer* **138**, 1528–1537 (2016).
- Cosgrove, S. E. *et al.* Initial low-dose gentamicin for *Staphylococcus aureus* bacteremia and endocarditis is nephrotoxic. *Clin. Infect. Dis.* **48**, 713–721 (2009).
- Álvarez, R., Vaz, B., Gronemeyer, H. & de Lera, A. R. Functions, therapeutic applications, and synthesis of retinoids and carotenoids. *Chem. Rev.* **114**, 1–125 (2014).

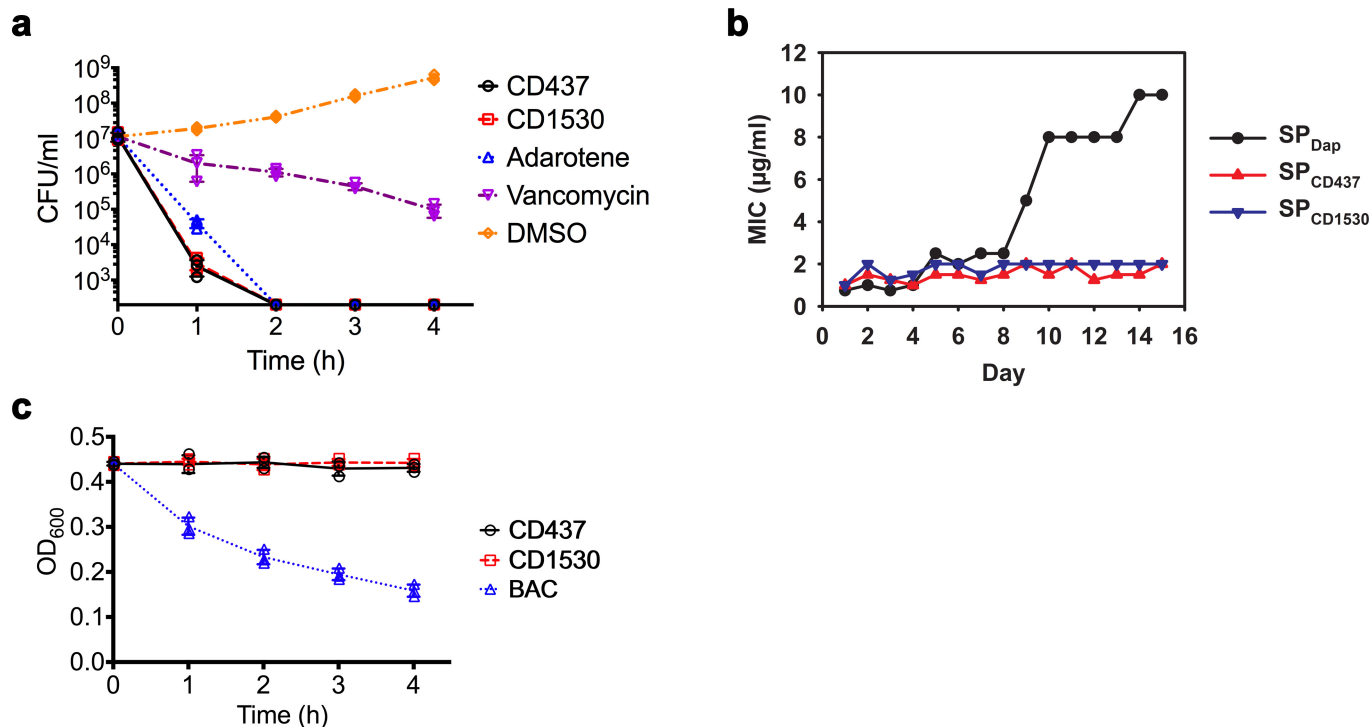
Supplementary Information is available in the online version of the paper.

Acknowledgements This study was supported by National Institutes of Health grant P01 AI083214 to M.S.G., F.M.A. and E.M., by National Science Foundation grant CMMI-1562904 to H.G., and by National Institute of General Medical Sciences grant 1R35GM119426 and National Science Foundation grant NSF1755698 to W.M.W. D.V.T. is supported by National Eye Institute grant EY028222. We thank the Institute of Chemistry and Cell Biology–Longwood at Harvard Medical School for providing the chemical libraries used in this study. We thank L. Rice for providing the *E. faecium* strains, K. Bayles and J. Endres for providing plasmid pBK123, J. Saavedra for assistance with next-generation sequencing library preparation, and S. Khalid for providing the atomic structures and force fields of the phosphatidylglycerol, Lys-PG and DPG lipids. The simulations reported were performed on resources provided by the Extreme Science and Engineering Discovery Environment through grant MSS090046 and the Center for Computation and Visualization at Brown University.

Author Contributions W.K., A.L.C., R.R., B.B.F., F.M.A. and E.M. designed the chemical screen. W.K., B.B.F. and R.R. performed the chemical screen. W.K. designed, performed and analysed MIC assays, dose–response *C. elegans* infection assays, membrane permeability assays, time-kill assays and transmission electron microscopy experiments. W.K. and D.V.T. designed, performed and analysed the selection of resistant mutants and whole genome sequencing. W.K., N.F. and P.M.V. designed, performed and analysed giant unilamellar vesicle experiments. W.K., W.Z. and H.G. designed, performed and analysed molecular dynamics simulations. A.D.S., C.E.K. and W.M.W. synthesized analogues. W.K. and B.B.F. designed, performed and analysed toxicity tests. W.K., G.L.H., S.S., W.P. and K.L. designed, performed and analysed animal studies. A.L.C., B.B.F., P.M.V., W.M.W., M.S.G., H.G., F.M.A. and E.M. contributed reagents, materials and/or analysis tools. E.M. supervised the project. W.K., W.Z., G.L.H., W.M.W., H.G., F.M.A. and E.M. wrote the manuscript.

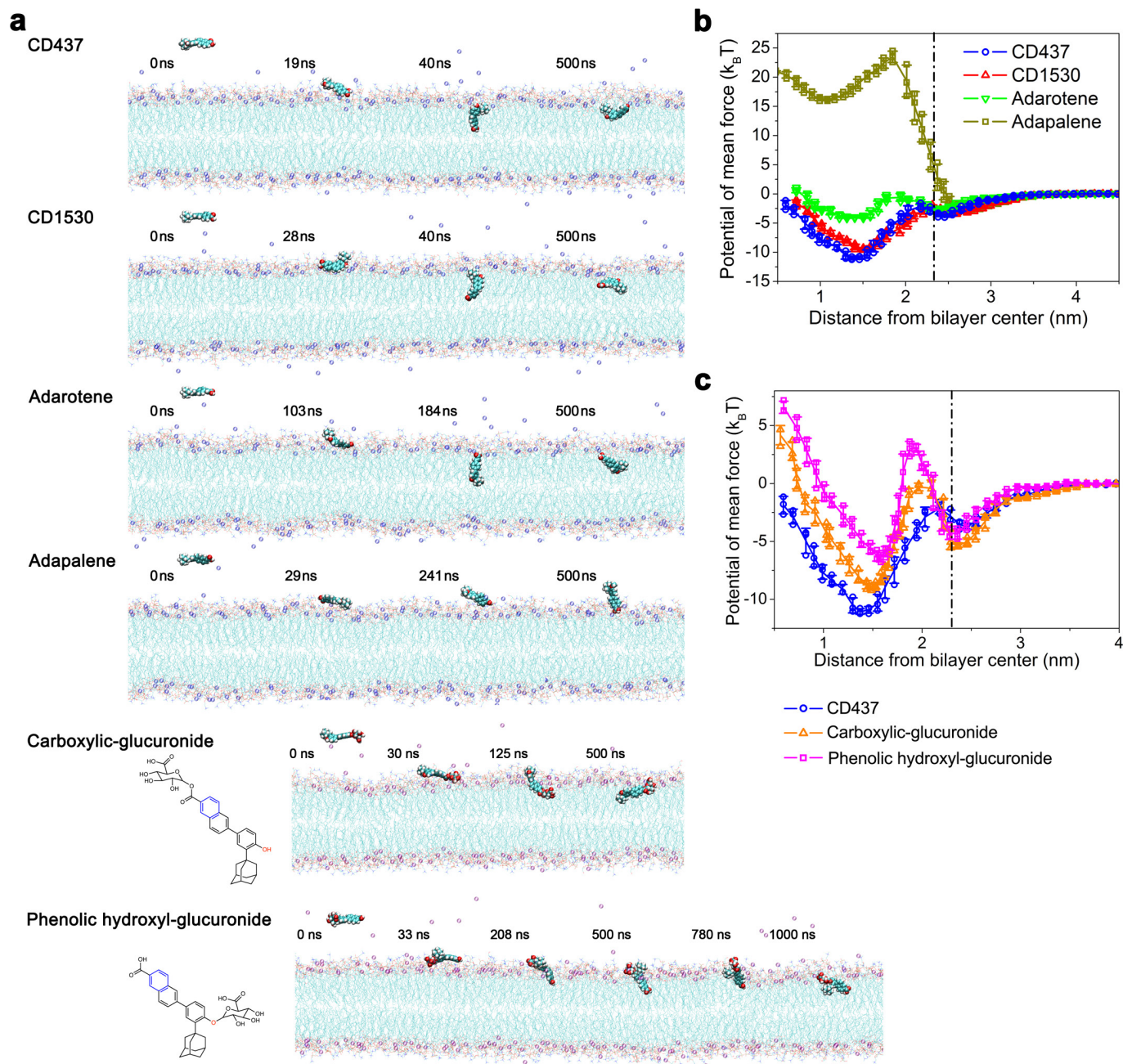
Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to E.M. (emylonakis@lifespan.org).

Reviewer Information Nature thanks F. DeLeo and the other anonymous reviewer(s) for their contribution to the peer review of this work.



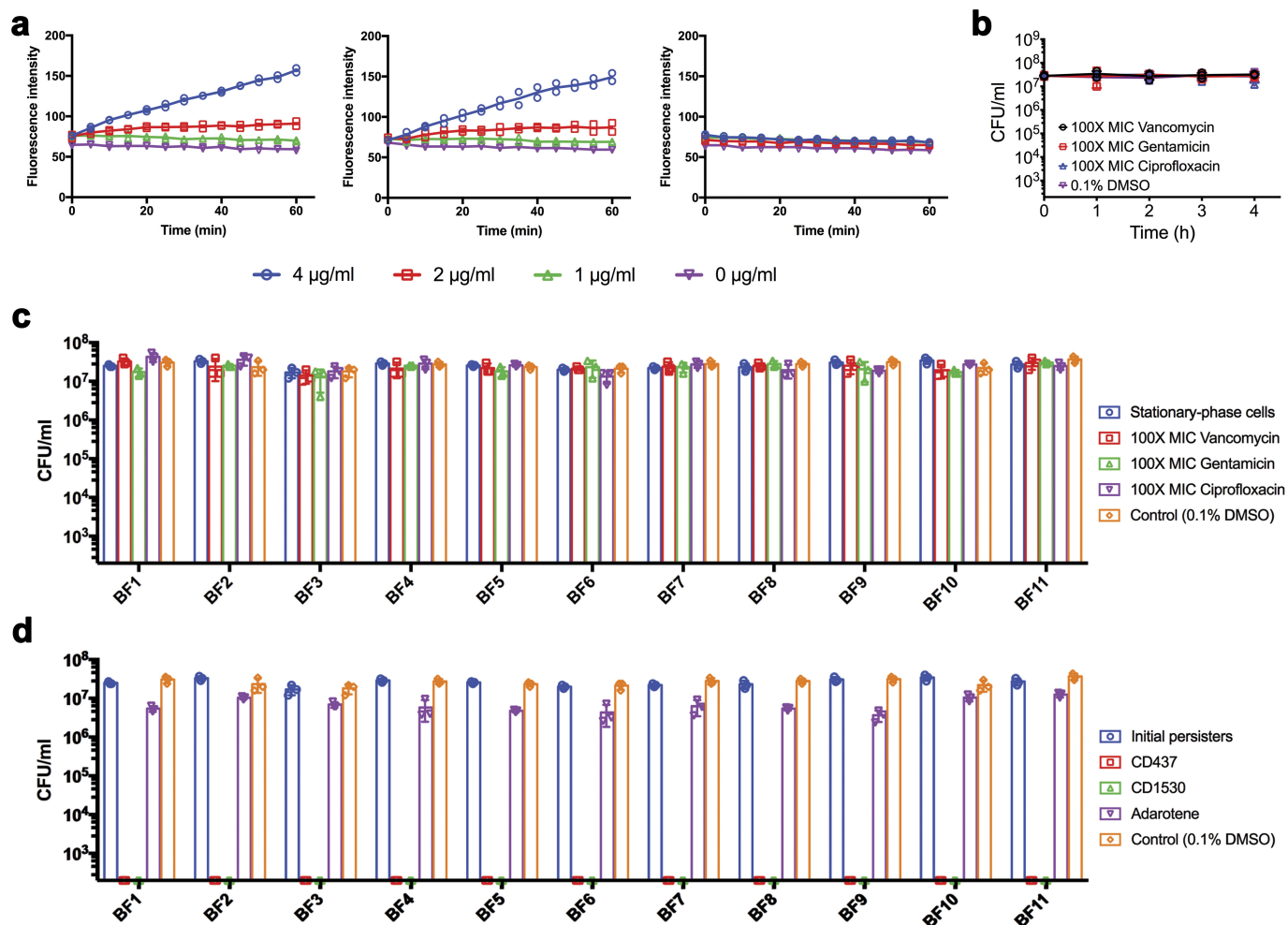
Extended Data Figure 1 | CD437 and CD1530 show fast-killing kinetics and low probability of resistance development, and do not cause detectable cell lysis. **a**, Exponential-phase MRSA cells (strain MW2) were treated with $10\times$ MIC CD437, CD1530, adarotene, vancomycin or 0.1% DMSO (negative control). CFU counts of cells were measured by serial dilution and plating on agar plates. The data points on the x axis are below the level of detection (2×10^2 CFU ml^{-1}). Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown.

b, Development of *S. aureus* MW2 mutants resistant to CD437 (SP_{CD437}), CD1530 ($\text{SP}_{\text{CD1530}}$) or daptomycin (SP_{Dap}) was attempted by daily serial passage for 15 days. **c**, Exponential-phase *S. aureus* MW2 bacteria were treated with $10\times$ MIC CD437, CD1530 or benzalkonium chloride (BAC) for 4 h. The anti-infective detergent BAC was used as a positive control for cell lysis. OD₆₀₀ was measured in a spectrophotometer every hour. Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown.



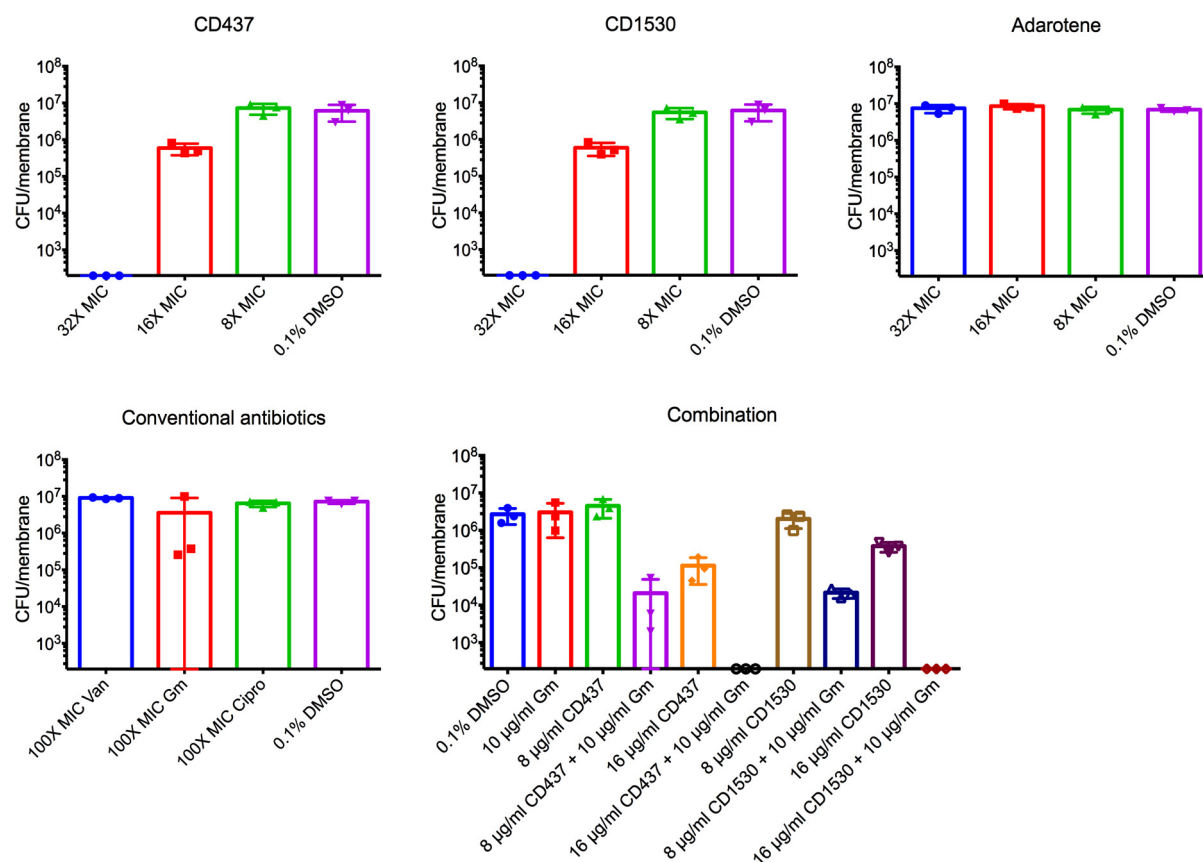
Extended Data Figure 2 | All-atom molecular dynamics simulations showing the interactions between selected retinoids or retinoid metabolites and a DOPC:DOPG (7:3) lipid bilayer. a, Representative configurations of synthetic retinoids or retinoid metabolites at, left to right, the onset of simulation, membrane attachment, membrane penetration and equilibrium state (see Supplementary Methods for atomic rendering). Simulations were repeated five times with similar results. **b, c,** Free energy profiles of the four retinoids (**b**) or CD437-metabolites (**c**) penetrating the membrane as a function of the

COM of the retinoids or the retinoid metabolites and the lipid bilayer. The dot-dashed line marks the membrane surface, averaged from the COM location of phosphate groups in the outer leaflet. Individual data points ($n = 3$ independent simulations) and mean \pm s.d. are shown. The membrane penetration of CD437, CD1530, adarotene, adapalene, the carboxylic-glucuronide metabolite and the phenolic hydroxyl-glucuronide metabolite are associated with transfer energies of $-8.92 k_B T$, $-7.14 k_B T$, $-1.45 k_B T$, $18.76 k_B T$, $-3.73 k_B T$, $-2.02 k_B T$ and energy barriers of $1.42 k_B T$, $1.12 k_B T$, $2.03 k_B T$, $26.13 k_B T$, $5.01 k_B T$, $7.40 k_B T$, respectively.



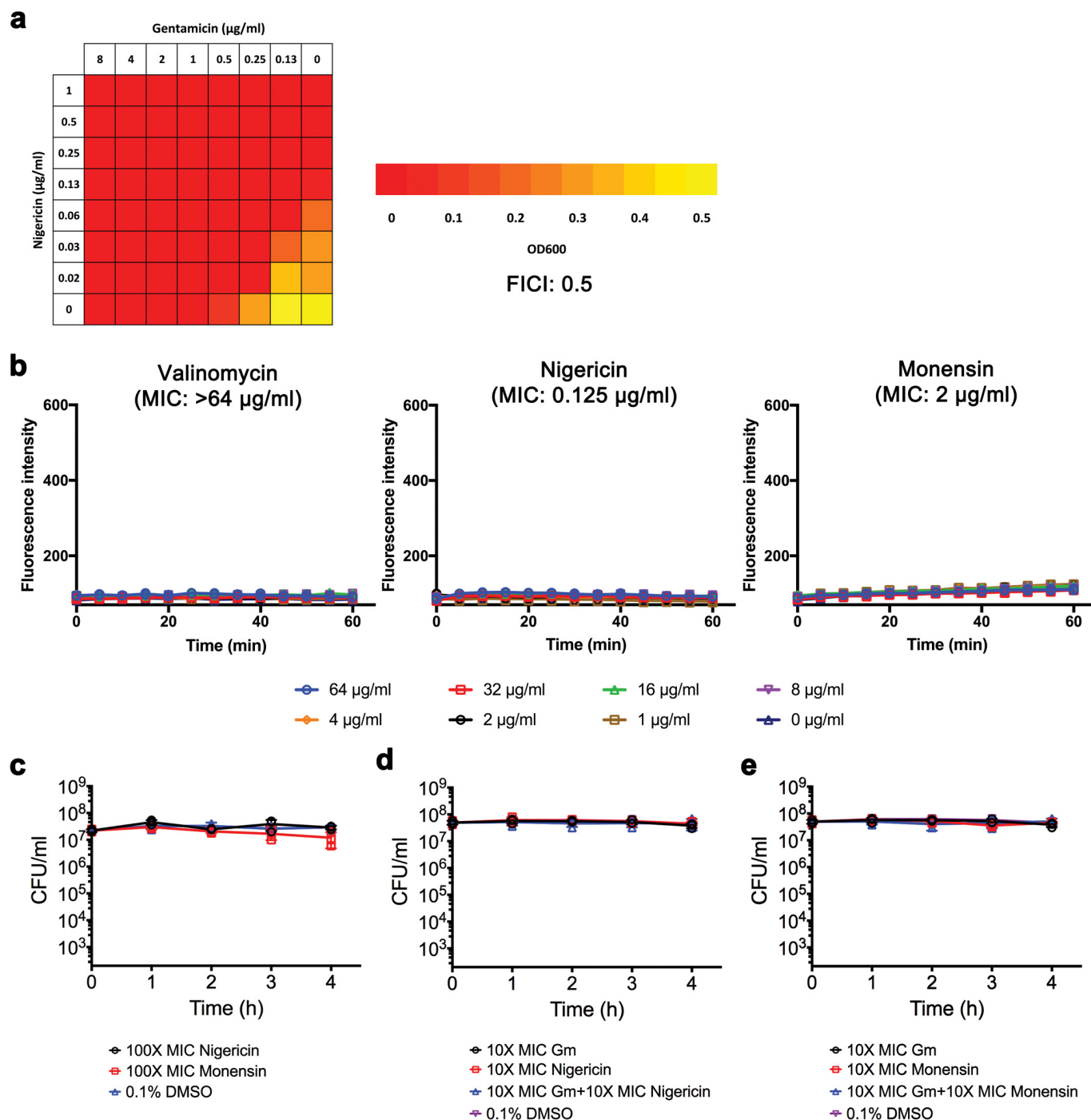
Extended Data Figure 3 | CD437 and CD1530 kill MRSA persisters by inducing membrane permeabilization. **a**, *S. aureus* MW2 persisters were treated with the indicated concentrations of the retinoids. Membrane permeability was measured spectrophotometrically by monitoring the uptake of SYTOX Green ($\lambda_{\text{ex}} = 485 \text{ nm}$, $\lambda_{\text{em}} = 525 \text{ nm}$) over time. Individual data points ($n = 2$ biologically independent samples) and means are shown; error bars are not shown for clarity. **b–d**, Stationary-phase

S. aureus MW2 (**b**) or stationary-phase cells of 11 clinical *S. aureus* isolates were treated with 100 \times MIC conventional antibiotics (**c**) or 10 \times MIC retinoids (**d**) for 4 h. Viability was measured by serial dilution and plating on agar plates. The data points on the x axis are below the level of detection ($2 \times 10^2 \text{ CFU ml}^{-1}$). **b–d**, Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown.



Extended Data Figure 4 | CD437 or CD1530 alone or in combination with gentamicin eliminate persisters formed in MRSA biofilms. MRSA MW2 biofilms formed on 13 mm cellulose ester membranes were treated with the indicated concentrations of retinoids alone or in combination

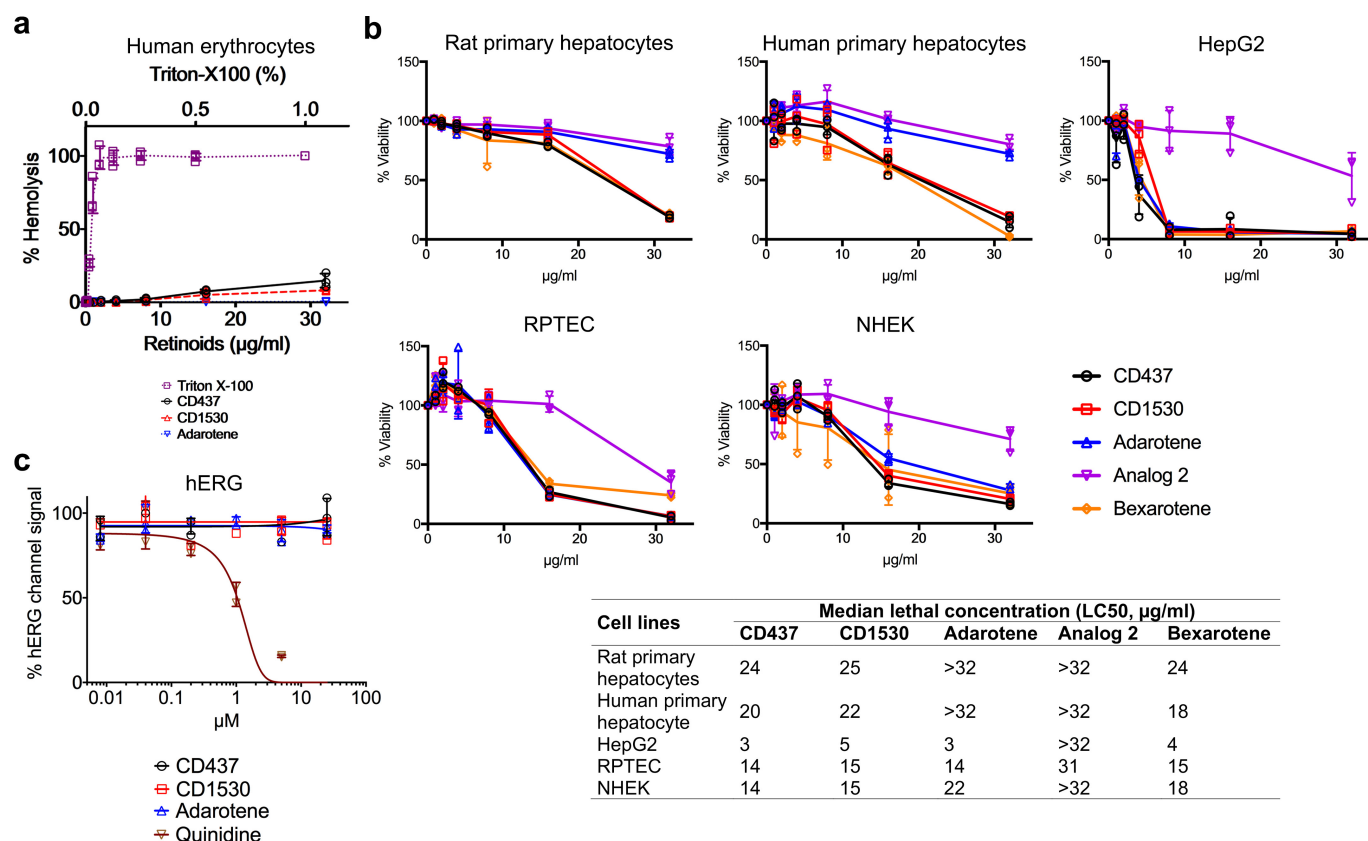
with gentamicin. The number of viable cells in biofilms was measured by CFU counting. The data points on the x axis are below the level of detection (2×10^2 CFU ml $^{-1}$). Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown.



Extended Data Figure 5 | Ionophores do not induce SYTOX Green membrane permeabilization or kill MRSA MW2 persisters.

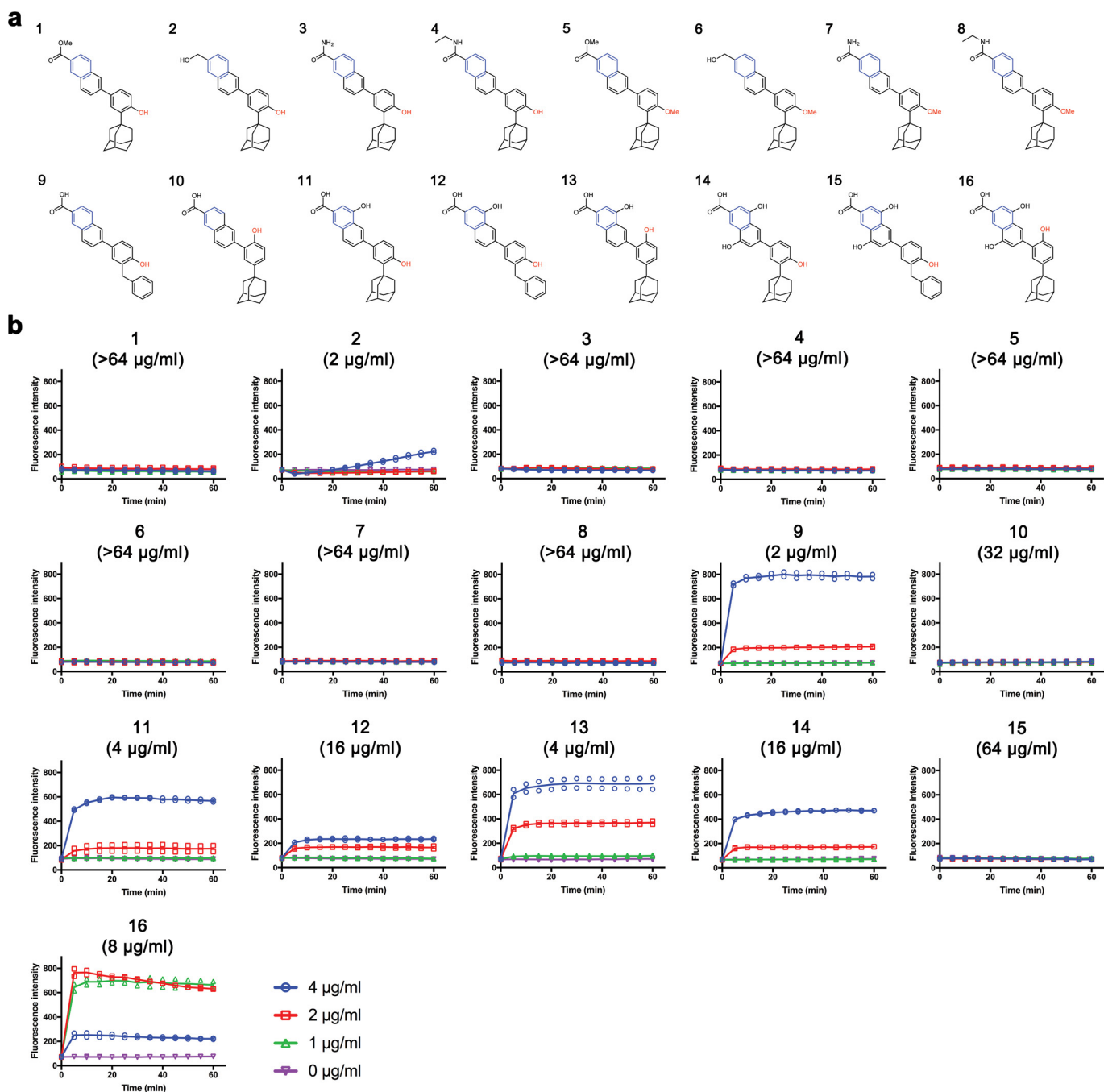
a, Synergism between nigericin and gentamicin was evaluated against *S. aureus* MW2 by the fractional inhibitory concentration index (FICI) microdilution checkerboard method. Optical densities at 600 nm were measured after 18 h incubation at 37 °C. Experiments were independently repeated twice with similar results. Synergy, $\text{FICI} \leq 0.5$; no interaction, $0.5 < \text{FICI} \leq 4$; antagonism, $\text{FICI} > 4$. **b**, Exponential-phase MW2 cells were treated with the indicated concentrations of valinomycin, nigericin or

monensin. Membrane permeability was measured spectrophotometrically by monitoring the uptake of SYTOX Green ($\lambda_{\text{ex}} = 485 \text{ nm}$, $\lambda_{\text{em}} = 525 \text{ nm}$) over time. Individual data points ($n = 2$ biologically independent samples) are shown; error bars are not shown for clarity. **c–e**, Stationary-phase *S. aureus* MW2 was treated with the indicated concentrations of ionophores, alone or combined with $10 \times$ MIC gentamicin (Gm), or 0.1% DMSO (control) for 4 h. Viability was measured by serial dilution and plating on agar plates. Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown.



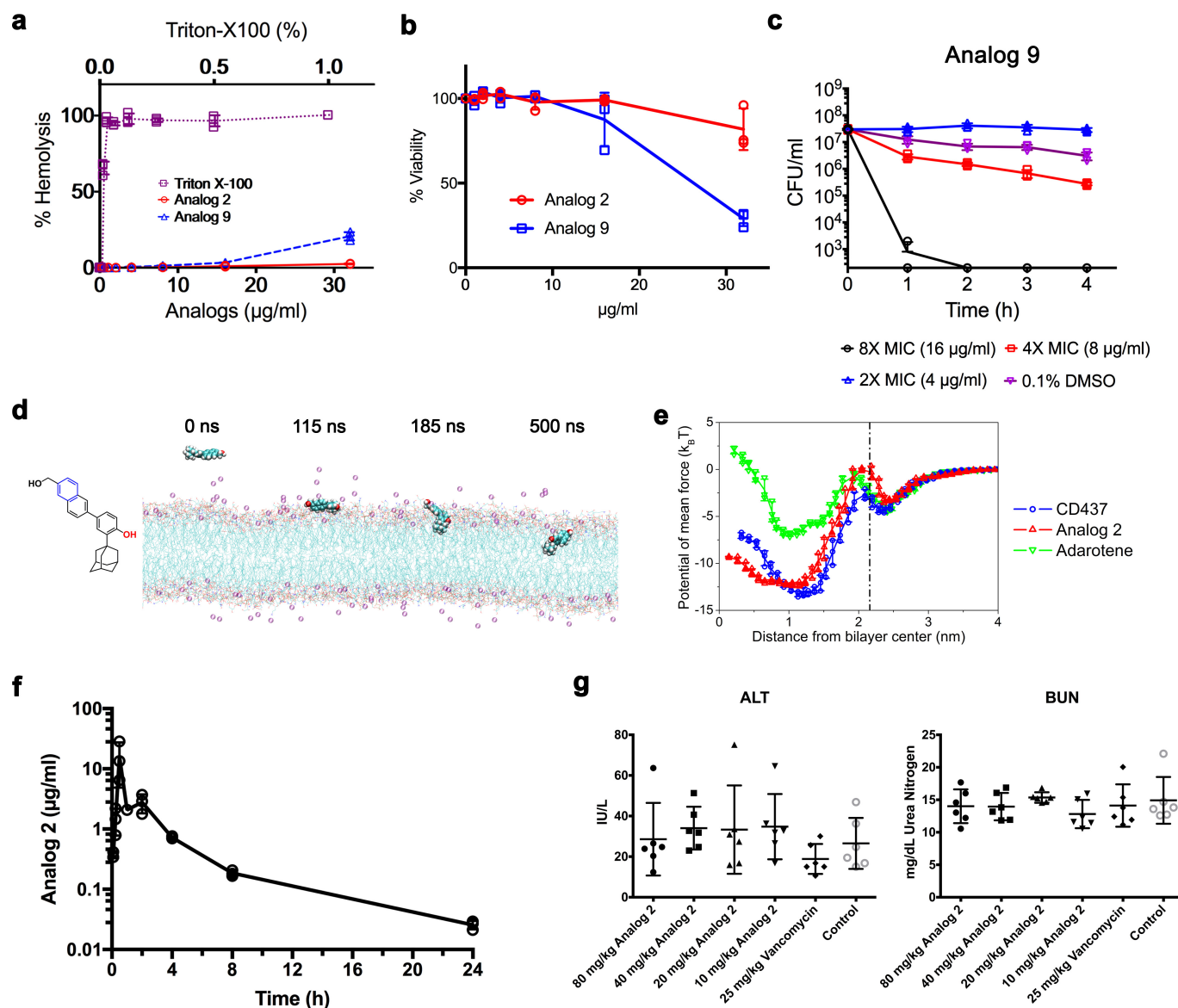
Extended Data Figure 6 | Evaluation of cytotoxic potentials of retinoids in various cell lines. **a**, Measurement of haemolytic activity. 2% human erythrocytes were treated with twofold serially diluted concentrations of the retinoids for 1 h at 37 °C. A sample treated with 1% Triton X-100 was used as the control for 100% haemolysis. **b**, Normal rat, human primary hepatocytes, human hepatoma (HepG2) cells, normal human primary renal proximal tubule epithelial cells (RPTEC) or adult normal human epidermal keratinocytes (NHEK) were treated with a range of concentrations of the synthetic retinoids in chemically defined, serum-free

medium for 24 h. The FDA-approved antineoplastic retinoid bexarotene was used as a control. Cell viability was calculated on the basis of absorbance readings at 450 nm at 4 h after adding WST-1. **a**, **b**, Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown. **c**, Three synthetic retinoids and the positive control quinidine were tested for inhibition of the hERG potassium channel. Individual data points ($n = 4$ biologically independent samples) and mean \pm s.d. are shown. Data are fitted to a standard inhibition curve.



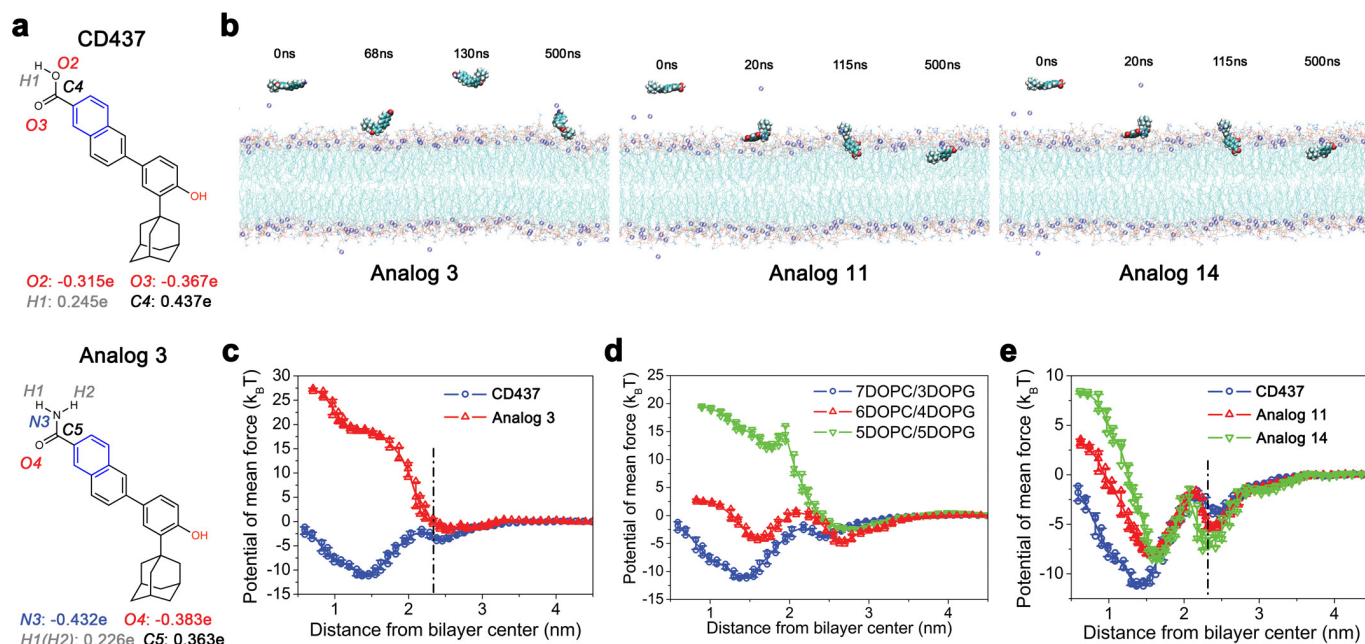
Extended Data Figure 7 | Structure-activity relationships. **a**, The chemical structures of newly synthesized CD437 analogues. **b**, MICs and membrane permeability were measured for *S. aureus* strain MW2. Membrane permeability was evaluated spectrophotometrically by

monitoring the uptake of SYTOX Green ($\lambda_{\text{ex}} = 485 \text{ nm}$, $\lambda_{\text{em}} = 525 \text{ nm}$) over time. Individual data points ($n = 2$ biologically independent samples) and means are shown; error bars are not shown for clarity.



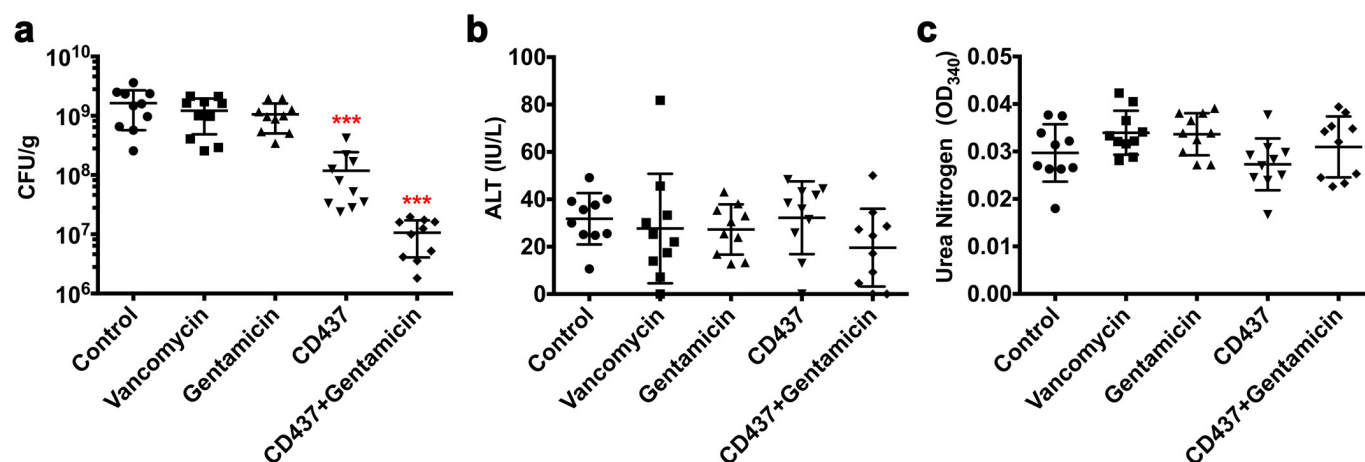
Extended Data Figure 8 | Determination of the biological properties of analogues 2 and 9. **a, b**, Human erythrocytes were treated for 1 h (**a**) and rat primary hepatocytes were treated for 24 h (**b**) with analogues 2 and 9. **c**, MRSA MW2 persisters were treated with analogue 9. The data points on the x axis are below the level of detection (2×10^2 CFU ml $^{-1}$). **a–c**, Individual data points ($n = 3$ biologically independent samples) and mean \pm s.d. are shown. **d**, Representative configurations of molecular dynamics simulations of analogue 2 interacting with lipid bilayers (108 phosphatidylglycerol lipids, 72 Lys-PG lipids and 10 DPG lipids; see Supplementary Methods for atomic rendering). Simulations were repeated five times with similar results. **e**, Free energy profiles of analogue 2, CD437 and adarotene penetrating the membrane as a function of the distance between the COM of the retinoids and the lipid bilayer. The dot-dashed line marks the membrane surface, averaged from the COM location of phosphate groups in outer leaflet. Individual data points ($n = 3$ independent simulations) and mean \pm s.d. are shown. **f**, The plasma concentrations of analogue 2 after a single injection of analogue 2 (20 mg kg $^{-1}$, i.p., 3 mice per time point) were measured

using LC–MS/MS. Pharmacokinetic analysis was conducted using Phoenix WinNonlin software version 6.3. Individual data points ($n = 3$ biologically independent animals) and mean \pm s.d. are shown. The determined pharmacokinetic parameters are T_{max} (the time taken to reach the maximum concentration) 0.5 h, C_{max} (maximum concentration observed) 16.14 µg ml $^{-1}$, AUC_{last} (area under the curve to last time point) 16.38 h·µg ml $^{-1}$, AUC_{inf} (area under the curve to infinite) 16.54 h·µg ml $^{-1}$, $t_{1/2}$ (half-life) 4.49 h, clearance 20.16 ml min $^{-1}$ kg $^{-1}$. **g**, Six mice per group ($n = 6$ biologically independent animals) were treated with control (5% Kolliphor + 5% ethanol, i.p.), vancomycin (25 mg kg $^{-1}$, i.p.) or analogue 2 (10–80 mg kg $^{-1}$, i.p.) every 12 h for 3 days. At 12 h after the last treatment, alanine aminotransferase (ALT) and blood urea nitrogen (BUN) were analysed. The concentrations of ALT (in international units per litre, IU l $^{-1}$) and BUN (mg dl $^{-1}$) in each mouse serum sample analysed are plotted as individual points and the mean \pm s.d. is shown. Control and antibiotic treatments were analysed by one-way ANOVA and post hoc Tukey test, which demonstrated a lack of significant differences ($P > 0.7$ for all ALT and BUN samples).



Extended Data Figure 9 | The charges and the number of branch groups affects membrane activity of CD437-like retinoids. **a**, Comparison of partial atomic charges between CD437 and analogue 3. **b**, Representative configurations of molecular dynamics simulations of analogues 3, 11, and 14 interacting with lipid bilayers (DOPC:DOPG, 7:3). The amide group in analogue 3 is repelled away from the membrane despite the attachment of the hydroxyl group. Atomic rendering is described in Supplementary Methods. Simulations were repeated five times with similar results.

c, d, Free energy profiles of analogue 3 penetrating DOPC:DOPG (7:3) lipid bilayers (**c**) and CD437 penetrating differently charged lipid bilayers (**d**). **e**, Analogues 11 and 14 penetrating DOPC:DOPG (7:3) lipid bilayers as a function of the distance between the COM of the retinoids and the lipid bilayer. The dot-dashed line marks the membrane surface, averaged from the COM location of phosphate groups in the outer leaflet. **c–e**, Individual data points ($n = 3$ independent simulations) and mean \pm s.d. are shown.



Extended Data Figure 10 | In vivo efficacy of CD437 alone or in combination with gentamicin in a deep-seated mouse thigh infection model. We chose a dose of 20 mg kg^{-1} CD437 to test its *in vivo* efficacy in the MRSA mouse deep-seated thigh infection model, because a dose of 20 mg kg^{-1} has shown *in vivo* efficacy in mouse xenograft cancer models^{27–29}. Ten MRSA MW2-infected mice per group ($n = 10$ biologically independent animals, see Supplementary Methods) were treated with control (5% Kolliphor + 5% ethanol, i.p.), vancomycin (25 mg kg^{-1} , i.p.), gentamicin (30 mg kg^{-1} , s.c.), CD437 (20 mg kg^{-1} , i.p.), or a combination of CD437 (20 mg kg^{-1} , i.p.) and gentamicin (30 mg kg^{-1} , s.c.) every 12 h

for 3 days beginning 24 h after infection. At 12 h after the last treatment, mice were euthanized and their thighs were excised and homogenized, and blood was collected and analysed for ALT and BUN. **a**, CFUs from each mouse thigh are plotted as individual points and the mean \pm s.d. for each experimental group is shown. **b**, **c**, Concentration of ALT for each mouse serum sample (**b**) and absorbance of BUN at 340 nm (**c**) are plotted as individual points. The mean \pm s.d. for each experimental group is shown. Statistical differences between control and antibiotic treatment groups were analysed by one-way ANOVA and post hoc Tukey test (***) $P < 0.0001$.

27. Schadendorf, D. *et al.* Treatment of melanoma cells with the synthetic retinoid CD437 induces apoptosis via activation of AP-1 *in vitro*, and causes growth inhibition in xenografts *in vivo*. *J. Cell Biol.* **135**, 1889–1898 (1996).
28. Langdon, S. P. *et al.* Growth-inhibitory effects of the synthetic retinoid CD437 against ovarian carcinoma models *in vitro* and *in vivo*. *Cancer Chemother. Pharmacol.* **42**, 429–432 (1998).

29. Ponzanelli, I. *et al.* Isolation and characterization of an acute promyelocytic leukemia cell line selectively resistant to the novel antileukemic and apoptogenic retinoid 6-[3-adamantyl-4-hydroxyphenyl]-2-naphthalene carboxylic acid. *Blood* **95**, 2672–2682 (2000).

Whole-organism clone tracing using single-cell sequencing

Anna Alemany^{1*}, Maria Florescu^{1*}, Chloé S. Baron^{1*}, Josi Peterson-Maduro^{1*} & Alexander van Oudenaarden¹

Embryonic development is a crucial period in the life of a multicellular organism, during which limited sets of embryonic progenitors produce all cells in the adult body. Determining which fate these progenitors acquire in adult tissues requires the simultaneous measurement of clonal history and cell identity at single-cell resolution, which has been a major challenge. Clonal history has traditionally been investigated by microscopically tracking cells during development^{1,2}, monitoring the heritable expression of genetically encoded fluorescent proteins³ and, more recently, using next-generation sequencing technologies that exploit somatic mutations⁴, microsatellite instability⁵, transposon tagging⁶, viral barcoding⁷, CRISPR–Cas9 genome editing^{8–13} and Cre–loxP recombination¹⁴. Single-cell transcriptomics¹⁵ provides a powerful platform for unbiased cell-type classification. Here we present ScarTrace, a single-cell sequencing strategy that enables the simultaneous quantification of clonal history and cell type for thousands of cells obtained from different organs of the adult zebrafish. Using ScarTrace, we show that a small set of multipotent embryonic progenitors generate all haematopoietic cells in the kidney marrow, and that many progenitors produce specific cell types in the eyes and brain. In addition, we study when embryonic progenitors commit to the left or right eye. ScarTrace reveals that epidermal and mesenchymal cells in the caudal fin arise from the same progenitors, and that osteoblast-restricted precursors can produce mesenchymal cells during regeneration. Furthermore, we identify resident immune cells in the fin with a distinct clonal origin from other blood cell types. We envision that similar approaches will have major applications in other experimental systems, in which the matching of embryonic clonal origin to adult cell type will ultimately allow reconstruction of how the adult body is built from a single cell.

The goal of our experiment is twofold: first, to link cells in the embryo to their corresponding clones in adult tissue (Fig. 1a); second, to quantify cell-type composition of these clones to determine the multipotency of embryonic progenitors. To reach the first goal, we need to uniquely label the cells in an embryo with permanent and heritable labels. For this, we use CRISPR–Cas9 technology, which induces a double-stranded break at the targeted genomic site that is repaired as insertions or deletions of different lengths at different positions (scars)^{16,17}. To allow for multiple scarring in the same cell, we use a zebrafish line with eight in-tandem copies of a histone–green fluorescent protein (GFP) transgene¹⁷ (Methods). Scarring starts after injecting the yolk or cell of the zygote with Cas9 RNA or protein, and a single-guide RNA (sgRNA) that targets GFP (Fig. 1b).

We quantified the scarring rate by measuring the fraction of unscarred GFP in zebrafish embryos at different times after Cas9 delivery (Fig. 1c), which is five times faster in Cas9 protein than in RNA injections. Cas9 activity ceases at around 3 h for protein and at 10 h for RNA injections, when zebrafish embryos have about 1,000 and 8,000 cells, respectively². We detect more than 1,000 distinct scars, the

abundances and probabilities of which span several orders of magnitude¹³ (Supplementary Information sections 1 and 2).

To detect scars and transcriptome from single cells, we developed ScarTrace, which integrates a nested PCR step after transcriptome conversion to cDNA into the sorting and robot-assisted transcriptome sequencing (SORT-seq) protocol¹⁸ (Fig. 1d). Because the histone–GFP transgene is transcribed, scars can be detected from mRNA and genomic DNA (gDNA). Detection from gDNA is preferred because GFP expression might be tissue specific, vulnerable to silencing and scars might affect the half-life of the mRNA. We assessed the efficiency of scar detection from mRNA and gDNA by comparing scar patterns of single cells from the caudal fin obtained using ScarTrace with and without reverse transcription (Fig. 1d, step 1). We detected 3.3 ± 0.3 (mean \pm s.e.m.) scars per clone on average, and approximately 25% of the cells remained unscarred and therefore do not contain clonal information (Extended Data Fig. 1a). Clone sizes from gDNA and gDNA–mRNA detection are very similar (Extended Data Fig. 1a–d), indicating that ScarTrace reliably detects scars from gDNA in single cells.

We next used ScarTrace to explore the clonal composition of haematopoietic cells isolated from the whole kidney marrow (WKM) of two protein-injected (P1 and P2) and two RNA-injected (R1 and R2) zebrafish. We found one and two major clones in P1 and P2, and eight and six in R1 and R2, respectively (Fig. 2a, b, Extended Data Fig. 2a, b, Extended Data Table 1). This is a direct result of the time window of Cas9 activity (Fig. 1c). The number of observed clones agrees with previous findings using GESTALT⁸, in which a similar Cas9-mediated approach is used to label embryonic clones in zebrafish, and with the number of clones (between 10.4 and 15.4) found for haematopoietic stem and progenitor cells at 10–14 hours post fertilization (hpf) using Zebrabow¹⁹.

The average number of scars per clone equals 3.3 ± 0.3 for P1, 1.02 ± 0.01 for P2, 3.5 ± 0.3 for R1 and 3.0 ± 0.3 for R2, with a minimum of 1 scar and a maximum of 5 scars per clone, revealing that both Cas9 protein and RNA efficiently cause scarring. We determined the copy number for each scar in a clone by modelling the amplification and sequencing noise of ScarTrace as a branching process (Supplementary Information section 3). Typically, the resulting number of scars per clone is smaller than eight, as a consequence of two or more simultaneously Cas9-induced cuts in the same multi-copy tandem histone–GFP gene¹⁰. We computed the *P* value of a combination of scars to occur in a cell (Fig. 2a, b). Values obtained are commonly below 10^{-6} , emphasizing that although identical scars might be independently introduced in different clones (for example, the yellow scar is present in one clone from fish R1 and four clones from fish R2), the chance of introducing the same combination of scars in independent clones is very small. Consistently, we do not find overlapping clones between different zebrafish. Using cell-to-cell variation in scar composition, we estimate a 90% scar detection efficiency (including unscarred GFP; Extended Data Fig. 1e, f). In addition, by assuming maximum parsimony for sequential scarring events, we build lineage trees for clones (Fig. 2c, d, Extended Data Fig. 2c, d, Supplementary Information section 4).

¹Onco Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center Utrecht, 3584 CT Utrecht, The Netherlands.

*These authors contributed equally to this work.

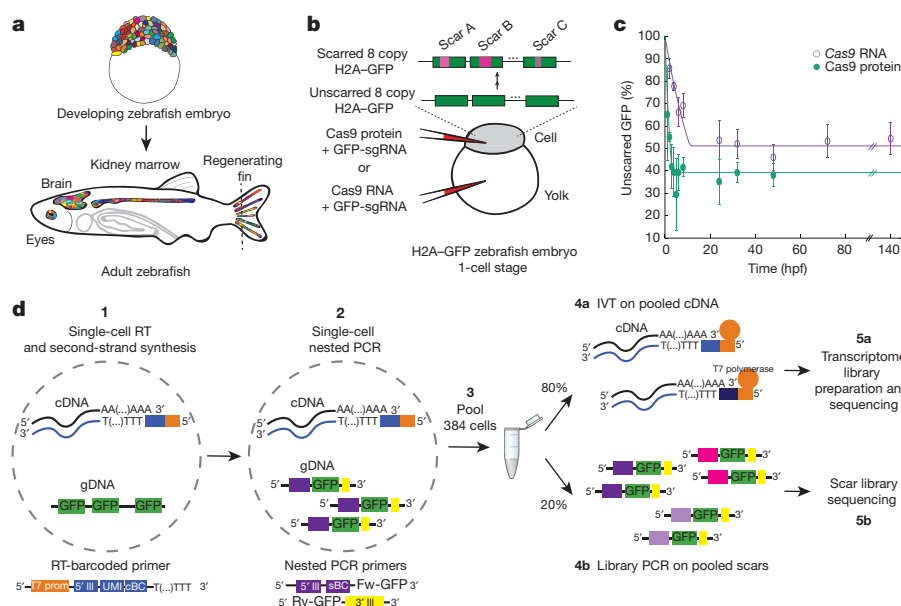


Figure 1 | Single-cell clonal tracing in zebrafish. **a**, Embryonic cells get permanent and unique labels that are transmitted to the clones in the adult. **b**, Zygote injection with Cas9 RNA or protein and sgRNA that targets GFP (GFP-sgRNA). H2A, histone 2A. **c**, Mean fraction of unscarred GFP as a function of time, computed over ten independently injected embryos ($t > 6$ h), and over three pools of ten embryos ($t \leq 6$ h), in which GFP was PCR-amplified from gDNA. Error bars denote s.e.m. Unscarred GFP exponentially decreases at $0.064 \pm 0.002 \text{ h}^{-1}$ ($0.294 \pm 0.008 \text{ h}^{-1}$) and is constant after 10 ± 1.0 h (3.1 ± 0.1 h) for RNA (protein) injections (solid lines, Supplementary Information section 1). **d**, ScarTrace protocol (Methods). IVT, *in vitro* transcription; RT, reverse transcription. cBC, cell-specific barcode for transcriptome; sBC, cell-specific barcode for scars.

Using RaceID²⁰ (Methods), we identify eight haematopoietic cell types in fish R1 and R2 (Fig. 2e). Gene expression profiles in the different cell types found for both fish are identical with the exception of erythrocytes, which show slight differences in the expression of characteristic markers (Extended Data Fig. 2e–h). After combining cell type and clonal information for single cells, we observe all clones in all cell types with similar proportions (Fig. 2f, g, Extended Data Fig. 2i, j), indicating that all clones contribute to the production of all blood cells. This is consistent with haematopoietic stem and progenitor cells specification (around 28 hpf), when scarring is already completed²¹.

Next, we used ScarTrace in the adult brain and eyes of two RNA-injected fish (R2 and R3), in which we identified different neuronal,

glia and immune cells (Fig. 3a, Extended Data Fig. 3). To determine clonal enrichment or depletion in certain cell types quantitatively, we used Fisher's exact test (Fig. 3b, Extended Data Fig. 4a). Here, several clones only generate neurons or retinal interneurons (Extended Data Fig. 4b, c). We observed that microglia share clones with the WKM, confirming that they originate from the WKM²².

Upon the exclusion of WKM clones, we found that clones are not only cell-type specific, but also brain-region and eye specific (Extended Data Figs 4d–g, 5a–d). Although R2 and R3 left and right midbrains share a small fraction of clones, left and right eyes share none (Fig. 3c, Extended Data Fig. 4h). However, for fish P1, both midbrains share almost all clones whereas eyes share only one. To explore when this

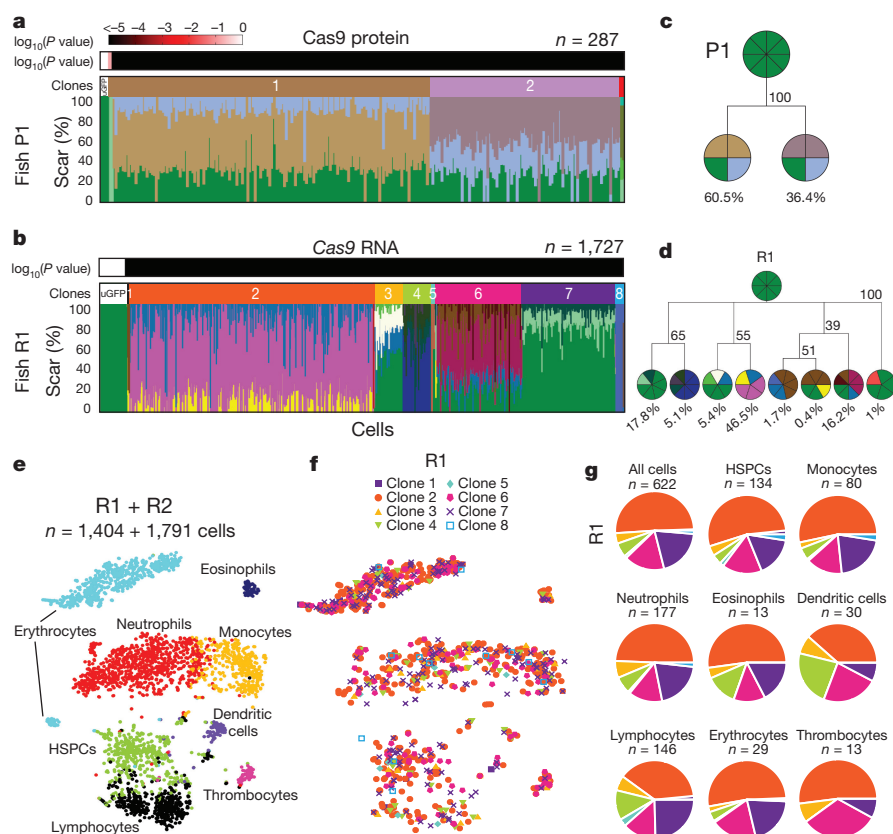


Figure 2 | Few clones produce haematopoietic cells. **a**, **b**, Scar percentage per cell for fish P1 (**a**) and R1 (**b**) (for key, see Extended Data Fig. 2a). The bar above each panel indicates clones and corresponding P values. uGFP, unscarred GFP. **c**, **d**, Lineage trees for clones in P1 (**c**) and R1 (**d**). The root is an unscarred clone. Clones with corresponding cell fraction and scar copy number (Supplementary Information section 3) are at the tips. The statistical confidence of each branch is computed as its proportion among 10,000 bootstrapped tree replicates. Only clones with more than 2 cells are taken into account. **e**, t -distributed stochastic neighbour embedding (t -SNE) map of cells from fish R1 and R2 obtained with transcriptome data. Colours indicate the cell type (Extended Data Fig. 2). HSPCs, haematopoietic stem and progenitor cells. **f**, t -SNE map of cells in fish R1. Colours indicate the clone. **g**, Clonal cell fraction per cell type for fish R1.

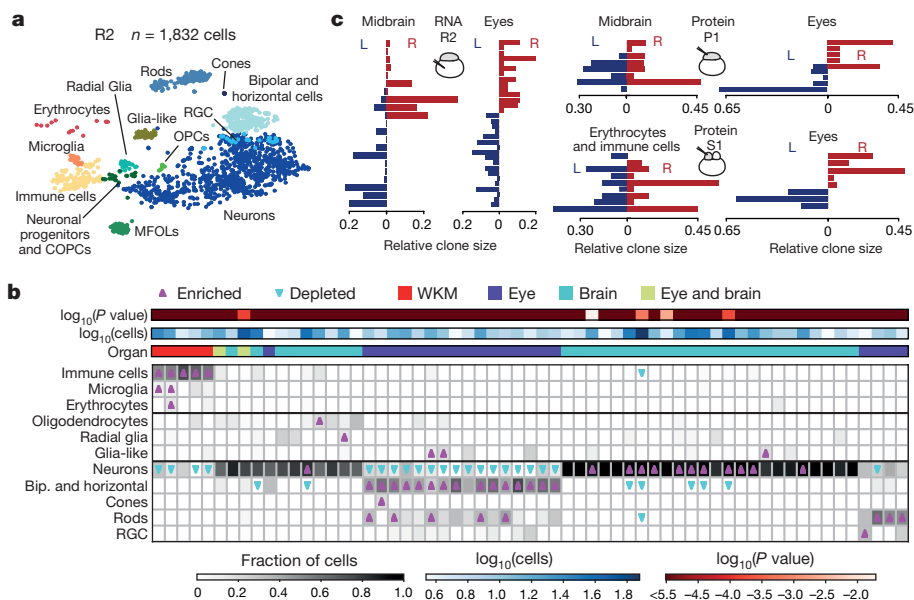


Figure 3 | Clonality in the brain and eyes.

a, *t*-SNE map of cells in fish R2. Colours indicate cell type (Extended Data Fig. 3). COPCs, committed oligodendrocyte progenitor cells; MFOLs, myelin-forming oligodendrocytes; OPCs, oligodendrocyte progenitors; RGC, retinal ganglion cell. **b**, Heat map of clonal cell fraction for cell types in fish R2 (COPC, OPC and MFOL clones merged as oligodendrocytes), and two-sided Fisher's exact test for enriched and depleted clones per cell type with $P < 0.05$. The bars at the top depict organ, total number of cells, and P value for each clone. Bip., bipolar; RGC, retinal ganglion cell. **c**, Relative clone frequency in the left (L) and right (R) eye and midbrain for fish R2 and P1, and left and right eye and erythrocytes and immune cells for fish S1.

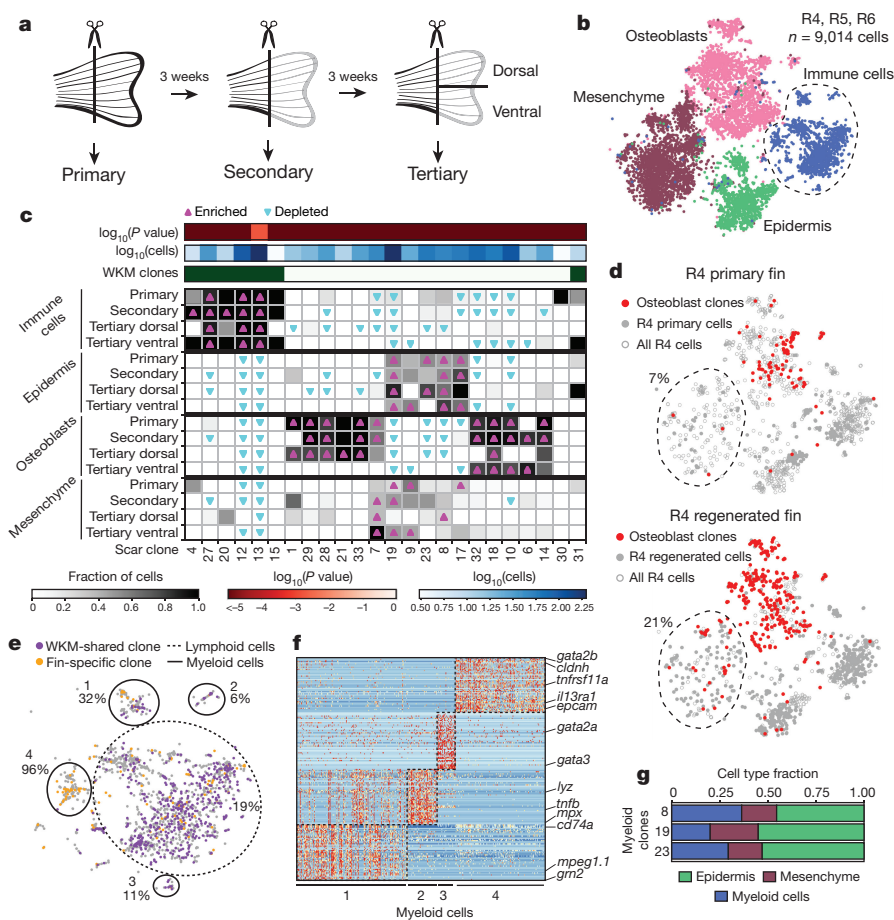


Figure 4 | Clonality during caudal fin regeneration. **a**, Primary (first week), secondary (fully regenerated after 3 weeks) and tertiary (fully regenerated after 6 weeks) fins are amputated. Tertiary fins are split dorsally and ventrally. **b**, *t*-SNE map of cells in fish R4, R5 and R6. Colours indicate cell type (Extended Data Fig. 6). **c**, Heat map of clonal cell fraction per cell type and fin in fish R4, and two-sided Fisher's exact test for clones that are enriched and depleted with $P < 0.05$. The bars at the top show clones in the WKM, number of cells and P value per clone. **d**, *t*-SNE map of cells from fish R4, with cells from the primary (top) or

regenerated (bottom) fin detected in osteoblast clones or remaining cell types. Percentages are mesenchymal cell fractions that share clones with osteoblasts. **e**, Magnified view of *t*-SNE map indicating immune cells with fin-specific clones, immune cells that share clones with the WKM and cells with no clone information (grey). Subpopulations of lymphoid cells and myeloid cells (numbered 1–4), and the percentage of RICs are indicated. **f**, Differential gene expression between the four subpopulations of myeloid cells (Supplementary Table 6). **g**, Cell type fraction for fin-specific clones detected in RICs in the primary fin for R4.

segregation is established, we injected one cell at the two-cell stage with Cas9–eScarlet fusion protein and sgRNA. We found Cas9–eScarlet protein present in only half of the embryo at dome stage (Extended Data Fig. 4i–l), approximately 3 h after Cas9 protein stops scarring. Therefore, scars only occur in one side of the embryo. However, ScarTrace on the left and right eyes of a 3-week-old-injected embryo (S1) reveals scars in both eyes. Upon removal of the clones found in immune cells and erythrocytes, the rest of the clones are specific to each eye (Fig. 3c). This indicates that both eyes get cellular contributions from both sides of the dome-stage embryo. To determine further when lateral commitment arises in eye progenitors, we built lineage trees for clones detected in the left and right eyes or midbrain for fish P1 and R2 (Extended Data Fig. 5e–h). In P1, no significant co-evolution is found among clones from the right (left) eye. By contrast, in R2 we observe a significant depletion of right eye clones evolving with left eye clones. This suggests that progenitors commit to the left or right eye shortly before the end of scarring with Cas9 protein. No significant co-evolution enhancement or depletion is found for clones detected in the left and right midbrain, indicating that cell mixing is important at 10 hpf. This is consistent with the processes of neurulation and neurogenesis²³.

Next, we focused on zebrafish caudal fin ontogeny and regeneration. We performed ScarTrace on the primary, secondary and tertiary fins of fish R4, R5 and R6 (Fig. 4a). We identified four major cell types (osteoblasts, mesenchymal, epidermal and immune cells) and observed cell-type-restricted clones in all fish (Fig. 4b, c, Extended Data Figs 6, 7a–e). We found that mesenchymal and epidermal cells share clones, revealing a common developmental origin that is maintained during regeneration. Together with previous imaging-based studies²⁴, this suggests that epidermal ancestors undergo epithelial-to-mesenchymal transition during gastrulation to generate mesenchymal cells in the caudal fin. Osteoblasts did not share clones with any other cell type in the primary fin and showed dorsal–ventral segregation, confirming their early lineage commitment during development^{13,25–28}. We found lineage restriction of the different cell types as the main mechanism of fin regeneration, consistent with previous results^{25,28}. However, after regeneration, we observed osteoblast-committed clones that generate a fraction (approximately 21% in R4, 44% in R6) of mesenchymal cells (Fig. 4d, Extended Data Fig. 7f). This suggests a certain degree of plasticity after injury, in which progenitors that produce osteoblasts during development can also give rise to mesenchymal cells during fin regeneration²⁹.

Finally, we investigated the clonal overlap of single cells from the WKM of fish R4, R5 and R6 with immune cells found in the fin. Clones detected in the WKM are enriched in the fin immune cells and depleted in the remaining cell types (Fig. 4c, e, Extended Data Fig. 7g). We found sub-populations of lymphoid and myeloid cells in all fins with different proportions of fin-specific clones, which we identify as resident immune cells (RICs). Differential gene expression analysis in myeloid cells revealed that subpopulation 4 expressed macrophage markers together with the epithelial marker *epcam* (Fig. 4f), which has been reported in resident macrophages in mice³⁰. All RICs in the primary fin share clonality with epidermal and mesenchymal cells (Fig. 4g, Extended Data Fig. 7b, c). Therefore, our data indicate that RICs have a distinct origin from haematopoietic stem cells (Extended Data Fig. 7h), and arise either from epidermal and mesenchymal transdifferentiation, or from ectodermal ancestors similarly to mesenchymal cells.

We developed ScarTrace as a new method to quantify clonal origin and cell type simultaneously at single-cell resolution. This enabled us to investigate the embryonic origin of clones found in different organs of the adult zebrafish and their cell-type commitment during development and regeneration. CRISPR–Cas9 genome editing technology for lineage tracing purposes at the single cell level has recently also been used in zebrafish to investigate lineages and cell types in the vertebrate brain, and to unravel developmental lineages^{31,32}. We anticipate many applications of ScarTrace in developmental and stem-cell biology, and similar approaches to study clonal selection in cancer models. Because ScarTrace provides a glimpse of the cellular past, it will be interesting to explore how this history is predictive of the current epigenetic and expression state.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 August 2017; accepted 2 February 2018.

Published online 28 March 2018.

1. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
2. Keller, P. J., Schmidt, A. D., Wittbrodt, J. & Stelzer, E. H. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science* **322**, 1065–1069 (2008).
3. Livet, J. et al. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56–62 (2007).
4. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
5. Reizel, Y. et al. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLoS Genet.* **7**, e1002192 (2011).
6. Sun, J. et al. Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
7. Naik, S. H. et al. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496**, 229–232 (2013).
8. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
9. Guernet, A. et al. CRISPR-barcoding for intratumor genetic heterogeneity modeling and functional analysis of oncogenic driver mutations. *Mol. Cell* **63**, 526–538 (2016).
10. Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. & Quake, S. R. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth. Biol.* **6**, 936–942 (2017).
11. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).
12. Frieda, K. L. et al. Synthetic recording and *in situ* readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
13. Junker, J. P. et al. Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars. Preprint at <https://www.biorxiv.org/content/early/2016/06/01/056499> (2016).
14. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized *in vivo*. *Nature* **548**, 456–460 (2017).
15. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
16. Jao, L. E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl Acad. Sci. USA* **110**, 13904–13909 (2013).
17. Pauls, S., Geldmacher-Voss, B. & Campos-Ortega, J. A. A zebrafish histone variant H2A.F/Z and a transgenic H2A.F/Z:GFP fusion protein for *in vivo* studies of embryonic development. *Dev. Genes Evol.* **211**, 603–610 (2001).
18. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394.e3 (2016).
19. Henninger, J. et al. Clonal fate mapping quantifies the number of haematopoietic stem cells that arise during development. *Nat. Cell Biol.* **19**, 17–27 (2017).
20. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
21. Jing, L. & Zon, L. I. Zebrafish as a model for normal and malignant hematopoiesis. *Dis. Model. Mech.* **4**, 433–438 (2011).
22. Xu, J. et al. Temporal-spatial resolution fate mapping reveals distinct origins for embryonic and adult microglia in zebrafish. *Dev. Cell* **34**, 632–641 (2015).
23. Schmidt, R., Strähle, U. & Scholpp, S. Neurogenesis in zebrafish — from embryo to adult. *Neural Dev.* **8**, 3 (2013).
24. Lee, R. T., Knapik, E. W., Thiery, J. P. & Carney, T. J. An exclusively mesodermal origin of fin mesenchyme demonstrates that zebrafish trunk neural crest does not generate ectomesenchyme. *Development* **140**, 2923–2932 (2013).
25. Tu, S. & Johnson, S. L. Fate restriction in the growing and regenerating zebrafish fin. *Dev. Cell* **20**, 725–732 (2011).
26. Knopf, F. et al. Bone regenerates via dedifferentiation of osteoblasts in the zebrafish fin. *Dev. Cell* **20**, 713–724 (2011).
27. Singh, S. P., Holdway, J. E. & Poss, K. D. Regeneration of amputated zebrafish fin rays from de novo osteoblasts. *Dev. Cell* **22**, 879–886 (2012).
28. Tornini, V. A. et al. Live monitoring of blastemal cell contributions during appendage regeneration. *Curr. Biol.* **26**, 2981–2991 (2016).
29. Tornini, V. A., Thompson, J. D., Allen, R. L. & Poss, K. D. Live fate-mapping of joint-associated fibroblasts visualizes expansion of cell contributions during zebrafish fin regeneration. *Development* **144**, 2889–2895 (2017).
30. Gautier, E. L. et al. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat. Immunol.* **13**, 1118–1128 (2012).
31. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* <http://doi.org/10.1038/nbt.4103> (2018).
32. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR/Cas9-induced genetic scars. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4124> (2018).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by a European Research Council Advanced grant (ERC-AdG 742225-IntScOmics), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) TOP award (NWO-CW 714.016.001), and the Foundation for Fundamental Research on Matter, financially supported by NWO (FOM-14NOISE01). This work is part of the Onco Institute which is partly financed by the Dutch Cancer Society. We thank M. Sen for help with sequencing, R. der Linden for cell sorting, and B. de Barbanson for help with programming, and all the other members of the A.v.O. laboratory for discussions and input. In addition, we thank B. Artegiani and J. Bakkers for discussions, P. Shang and N. Geijsen for sharing the Cas9–eScarlet fusion protein, the Hubrecht Sorting Facility, and the Utrecht Sequencing Facility, subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University.

Author Contributions A.v.O. conceived and designed the project. J.P.-M. developed the experimental protocol, with support from A.A., M.F. and C.S.B.

C.S.B. performed WKM-related experiments; M.F. performed brain- and eye-related experiments; and C.S.B. and J.P.-M. performed fin-related experiments. A.A. developed the computational methods and modelling. A.A., C.S.B. and A.v.O. analysed WKM-related data; A.A. and M.F. analysed brain- and eye-related data; A.A., C.S.B. and J.P.-M. analysed fin-related data. All authors discussed and interpreted results, and wrote the manuscript. A.A., M.F., C.S.B. and J.P.-M. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu).

Reviewer Information *Nature* thanks L. Zon and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Zebrafish Cas9 and sgRNA injections. Heterozygous zygotes of the transgenic zebrafish line Tg(h2afva:GFP)^{kca66};(h2afva:GFP)^{kca66} (ref. 17) were injected at the cell with 1 nl Cas9 protein (NEB; final concentration 1,590 ng μl⁻¹) or at the yolk with 1 nl Cas9 RNA (300 ng μl⁻¹) in combination with an sgRNA that targets GFP (25 ng μl⁻¹, sequence: GGTGTTCTGCTGGTAGTGGT) (Fig. 1b). Cas9 RNA was *in vitro* transcribed from a linearized pCS2-nCas9n vector (Addgene plasmid 47929)¹⁶ using the mMESAGE mMACHINE SP6 Transcription Kit (Thermo Scientific). The sgRNA was *in vitro* transcribed from a template using the MEGAscript T7 Transcription Kit (Thermo Scientific). The sgRNA template was synthesized with T4 DNA polymerase (New England Biolabs) by partially annealing two single-stranded DNA oligonucleotides containing the T7 promoter and the GFP-binding sequence, and the tracrRNA sequence, respectively⁵⁶. Male and female zebrafish were used, no randomization was done, no blinding was done and no animals were excluded from the analysis. No statistical methods were used to predetermine sample size. The age of the fish used in isolated organ spans 3–18 months. For sample sizes, see Extended Data Table 2. All animal experiments were performed in accordance with institutional and governmental regulations, and were approved by the Dier Experimenten Commissie of the Royal Netherlands Academy of Arts and Sciences and performed according to the guidelines.

Transgene copy number. To determine the number of integrations of the transgene, we performed whole-genome sequencing (NEBNext Ultra library preparation kit for Illumina (E7370S) and the NEB Multiplex Oligos for Illumina (E7500L)) on an homozygous Tg(h2afva:GFP)^{kca66};(h2afva:GFP)^{kca66} fish. Paired-end data were trimmed (TrimGalore-0.4.3) and mapped (bwa-0.7.10 mem) to the zebrafish reference genome (danRer10 from UCSC Genome Browser), and PCR and optical duplicates were removed (Picard-2.0.1) (Extended Data Fig. 8a, b). The copy number was extracted using FREEC-11.0⁵⁷ with default parameters. With a 1-kb window size, we find 19 copies of the transgene fragment, whereas with a 500-bp window size, we find 18 (Extended Data Fig. 8b). After correcting for reads due to endogenous copies, we estimate the number of copies of the transgene in a heterozygous fish to be 8 ± 1. This number agrees with single-cell data, because although we detected a maximum of 7 scars per clone (Extended Data Fig. 8c, d), we see that sometimes 6 of the scars in those clones represent approximately 12.5% of the scar content per cell, and one represents approximately 25% (Extended Data Fig. 8e). This again suggests that the number of integrations of the histone-GFP transgene is 1/0.125 = 8.

ScarTrace protocol. Live single cells (based on DAPI exclusion and scatter properties) were sorted into 384-well plates (Sigma-Aldrich) containing 5 μl of mineral oil (Sigma-Aldrich), 50 nl of uniquely barcoded reverse transcription primers (Supplementary Table 1), dNTPs (Promega), Spike-in controls (Thermo Fisher) and RNase inhibitor (SUPERaseIn, Thermo Fisher). Plates were immediately spun down and stored at -80 °C. Cells were lysed at 65 °C for 5 min. Reverse transcription and second-strand synthesis mixes were dispensed into each well using the Nanodrop II and reactions were performed at 42 and 16 °C degrees, respectively (Fig. 1d, step 1). Genomic DNA was access by proteinase K treatment followed by a nested PCR strategy to amplify the scarred GFP region (Fig. 1d, step 2). In the second PCR, unique scar barcodes were introduced in each well (Supplementary Table 2). All cells were pooled and the aqueous phase was separated from the oil phase (Fig. 1d, step 3). The collected material was split for scar library and transcriptome library preparation (Fig. 1d, step 4). For transcriptome library preparation, the SORT-seq protocol¹⁸ was used (Fig. 1d, step 5a). For scar library preparation, a PCR introducing only Illumina TruSeq adapters was performed (Fig. 1d, step 5b). All libraries were sequenced paired-end at 75 bp read-length on the Illumina NextSeq platform. A detailed description of the protocol is available in Protocol Exchange³³.

WKM isolation. The WKM was isolated as previously described³⁴. A ventral midline incision was made to open the adult zebrafish body cavity. All internal organs were carefully removed to access the kidney. The WKM was collected in PBS supplemented with FCS. The tissue was aspirated through a 1 ml pipet tip several times to mechanically dissociate haematopoietic cells. After two consecutive filtering steps (using 70-μm and 40-μm cell strainers (VWR)), cells were centrifuged and washed. The pellet of haematopoietic cells was resuspended in PBS and FCS supplemented with DAPI (Thermo Fisher) to assess cell viability.

Brain parts and eye isolation. Brain and eyes were isolated from the zebrafish head and dissected in PBS. Optic nerves were removed. The forebrain (olfactory bulb and telencephalon) was isolated from the midbrain, followed by dissection of the hindbrain (rhombencephalon). The midbrain (mesencephalon) was dissected into left and right midbrain. The eyes lens was carefully removed. Brain parts and eyes were dissociated into single cells using a papain-based solution (Thermo Fisher, 88285) and washing solutions as previously described³⁵. The washed cell pellet was resuspended in DMEM/F12 medium (Thermo Fisher, 11320033) and supplemented with DAPI (Thermo Fisher) to assess cell viability for FACS.

Fin amputation. Caudal fin amputations were performed as previously described³⁶, after which fish were returned to 28 °C aquarium water. Once isolated, this tissue was immediately dissociated by moderately shaking at 30 °C for 1 h, with gentle trituration performed every 10 min with a p200 pipet, in a solution of 2 mg ml⁻¹ collagenase A (Sigma-Aldrich) and 0.3 mg ml⁻¹ protease (type XIV, Sigma-Aldrich) in Hanks solution. After 1 h, the solution was incubated for 5 min in 0.05% trypsin in PBS. The solution was strained using 70-μm and 40-μm cell strainers (Corning) and cells were washed in 2% FBS in Hanks solution. Before flow cytometry, cells were centrifuged and resuspended in PBS and FBS supplemented with DAPI (Thermo Fisher) to assess cell viability.

Transcriptome analysis. In transcriptome libraries, the first read contains cell barcode (Supplementary Table 1) and unique molecular identifier (UMI) information, and the second read contains biological information. Second reads with a valid cell barcode extracted from corresponding first reads are mapped using bwa mem-0.7.10 with default parameters to the reference zebrafish transcriptome (*Danio rerio* assembly Zv9, ensemble 74, extended with ERCC92). For each cell, the number of transcripts per gene was obtained as previously described³⁷. We refer to transcripts as unique molecules based on UMI correction. We ran RaceID3 with different parameters for each organ under study (Supplementary Data 1) for cell filtering, normalization, gene filtering, cell clustering and differential gene expression analysis (in which *P* values are calculated using negative binomial distribution and corrected for multiple testing by the Benjamini–Hochberg method). The choice of filtering parameters was made to include the maximum number of cells in our analysis without losing cell type information. Supplementary Tables 3–6 provide results for the differentially expressed genes for each cell type compared with all other cells in the organ: WKM^{38–41} (90 dendritic cells, 76 eosinophils, 641 erythrocytes, 516 haematopoietic stem and progenitor cells, 446 lymphocytes, 409 monocytes, 927 neutrophils and 76 thrombocytes), brain and eyes^{42–51} (250 bipolar and horizontal cells, 45 COPCs, 9 cones, 290 erythrocytes, 254 immune cells, 88 glia-like cells, 89 MFOLs, 66 microglia, 1,427 neurons, 10 OPCs, 31 RCL, 53 radial glia and 202 rods), caudal fin^{52–54} (144 epidermal cells, 2,834 fibroblasts, 1,784 immune cells, and 2,951 osteoblasts), and resident myeloid cell types in the fin (118 cells in subpopulation 1, 45 in subpopulation 2, 27 in subpopulation 3 and 133 in subpopulation 4).

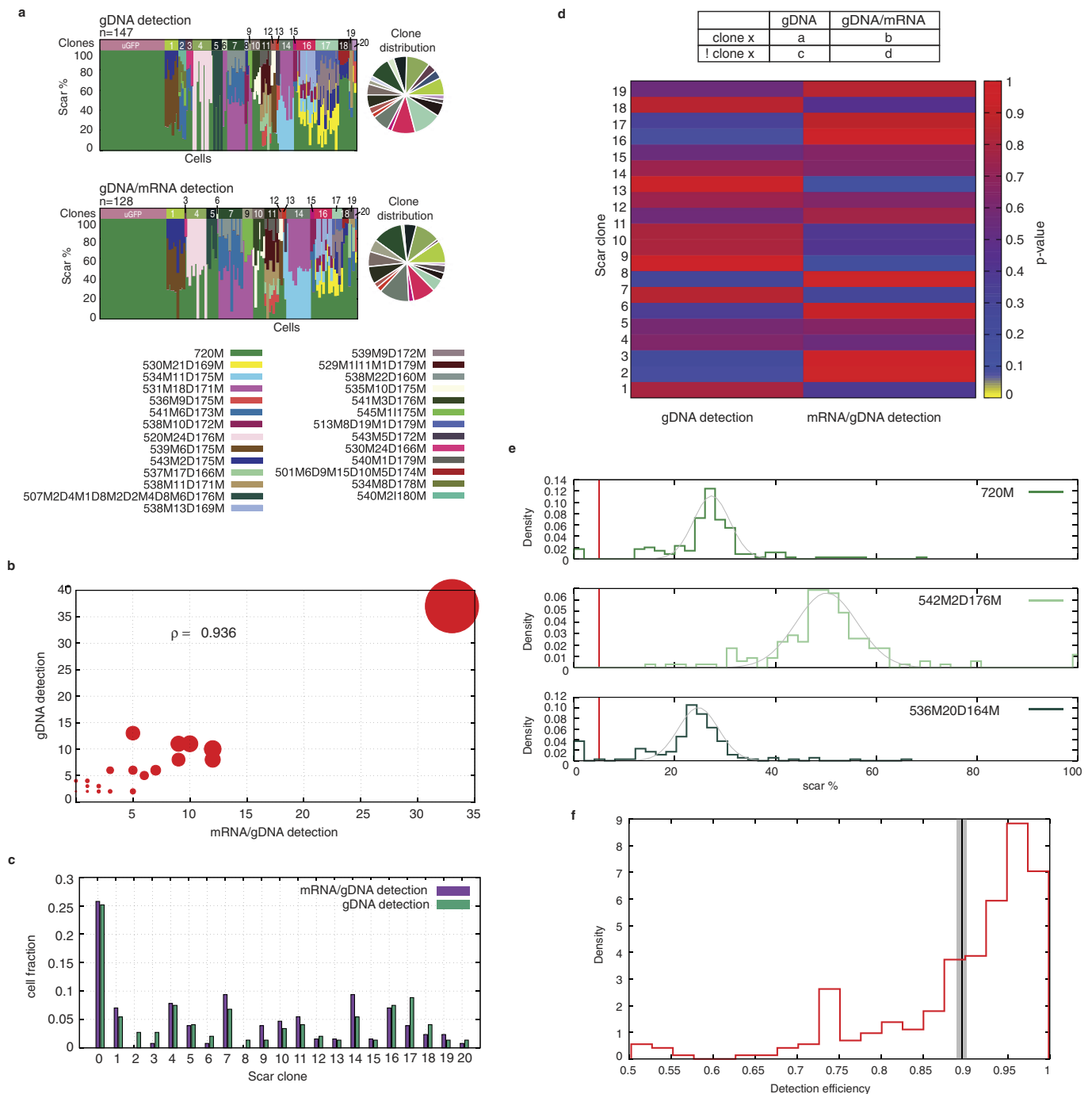
Scar analysis. In scar libraries, the first read contains the cell barcode (Supplementary Table 2) and the forward primer used in the nested PCR and second read contains the sequence for the scar and the reversed primer. Scripts to extract scars and detect clones are provided as Supplementary Data 2, together with a reference manual (Supplementary Data 3). Bug fixes and updates of the scripts can be downloaded from <https://github.com/anna-alemany/scScarTrace>. Cells sharing an identical scar pattern are assumed to come from the same clone, independently of scar percentage. Cells with a detected scar pattern that can be assigned to another single clone by assuming that some scar was not sampled were pooled with that clone. Cells that according to their scar pattern can be ambiguously assigned to two or more other clones were removed from subsequent analysis. Clones with less than three cells were also removed.

Code availability. Transcriptome analysis was performed using RaceID3 available at https://github.com/dgrun/RaceID3_StemID2, with parameters summarized in Supplementary Data 1. Scripts for scar extraction and clone detection are provided in Supplementary Data 2, together with a reference manual (Supplementary Data 3). Bug fixes and updates of the scripts can be downloaded from <https://github.com/anna-alemany/scScarTrace>.

Data availability. The accession numbers for the RNA sequencing datasets reported in this paper have been deposited with the Gene Expression Omnibus (GEO) under accession GSE102990.

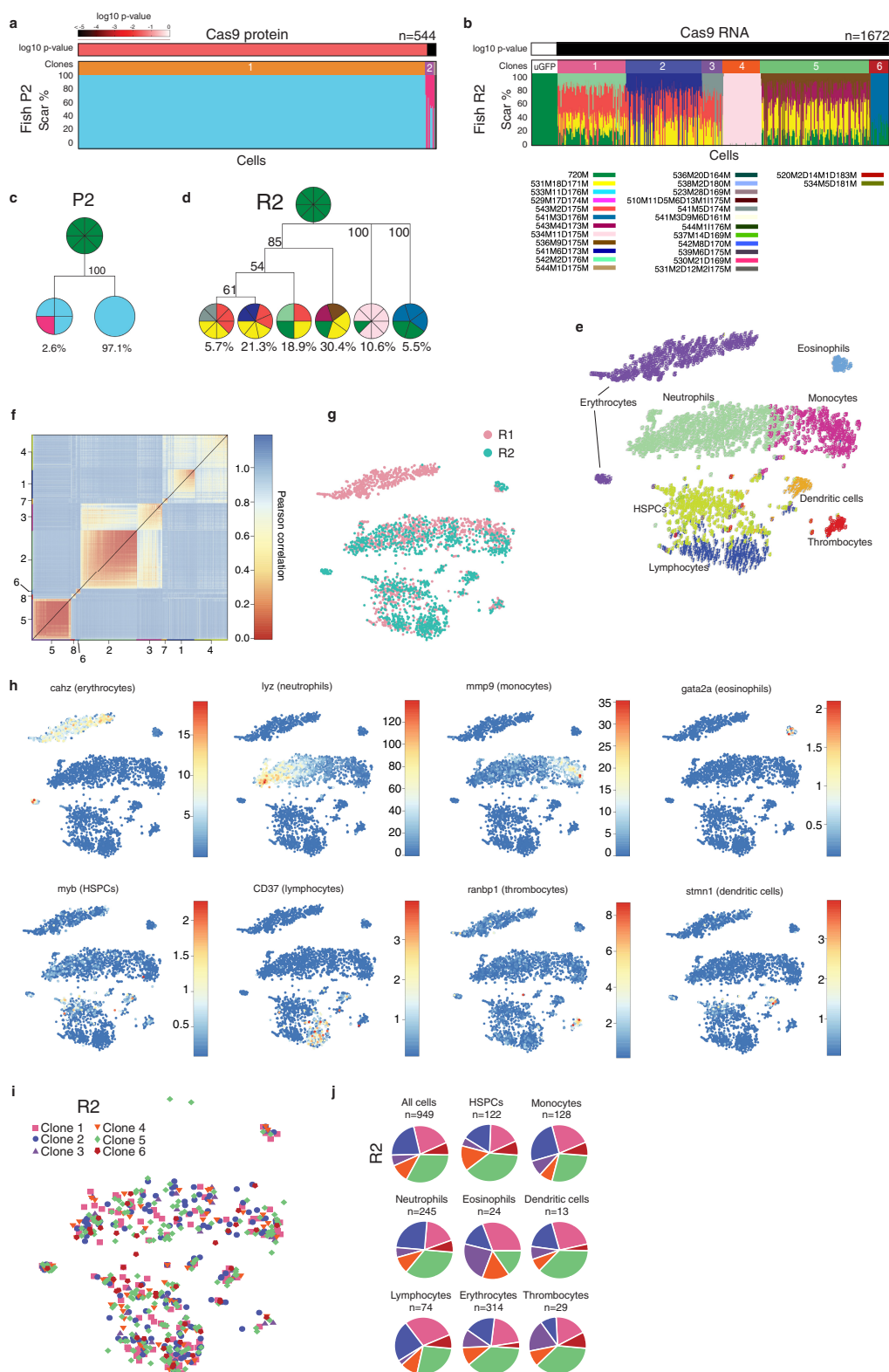
33. Peterson-Maduro, J., Florescu, M., Baron, C. S., Alemany, A. & van Oudenaarden, A. Single-cell ScarTrace. *Protoc. Exch.* <https://dx.doi.org/10.1038/protex.2018.017> (2018).
34. Stachura, D. L. & Traver, D. Cellular dissection of zebrafish hematopoiesis. *Methods Cell Biol.* **101**, 75–110 (2011).
35. Lopez-Ramirez, M. A., Calvo, C. F., Ristori, E., Thomas, J. L. & Nicoli, S. Isolation and culture of adult zebrafish brain-derived neurospheres. *J. Vis. Exp.* **108**, 53617 (2016).
36. Poss, K. D. et al. Roles for Fgf signaling during zebrafish fin regeneration. *Dev. Biol.* **222**, 347–358 (2000).
37. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
38. Kobayashi, I. et al. Comparative gene expression analysis of zebrafish and mammals identifies common regulators in hematopoietic stem cells. *Blood* **115**, e1–e9 (2010).
39. Moore, F. E. et al. Single-cell transcriptional analysis of normal, aberrant, and malignant hematopoiesis in zebrafish. *J. Exp. Med.* **213**, 979–992 (2016).
40. Macaulay, I. C. et al. Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Reports* **14**, 966–977 (2016).

41. Carmona, S. J. *et al.* Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. *Genome Res.* **27**, 451–461 (2017).
42. Cahoy, J. D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
43. Nelson, S. M., Frey, R. A., Wardwell, S. L. & Stenkamp, D. L. The developmental sequence of gene expression within the rod photoreceptor lineage in embryonic zebrafish. *Dev. Dyn.* **237**, 2903–2917 (2008).
44. Zhang, H., Copara, M. & Ekstrom, A. D. Differential recruitment of brain networks following route and cartographic map learning of spatial environments. *PLoS ONE* **7**, e44886 (2012).
45. Hickman, S. E. *et al.* The microglial sensome revealed by direct RNA sequencing. *Nat. Neurosci.* **16**, 1896–1905 (2013).
46. Di Donato, V., Auer, T. O., Duroure, K. & Del Bene, F. Characterization of the calcium binding protein family in zebrafish. *PLoS ONE* **8**, e53299 (2013).
47. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, D996–D1008 (2013).
48. La Manno, G. *et al.* Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).
49. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326–1329 (2016).
50. Chen, R., Wu, X., Jiang, L. & Zhang, Y. Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Reports* **18**, 3227–3241 (2017).
51. Oosterhof, N. *et al.* Identification of a conserved and acute neurodegeneration-specific microglial transcriptome in the zebrafish. *Glia* **65**, 138–149 (2017).
52. Marie, P. J. Transcription factors controlling osteoblastogenesis. *Arch. Biochem. Biophys.* **473**, 98–105 (2008).
53. Akerberg, A. A., Stewart, S. & Stankunas, K. Spatial and temporal control of transgene expression in zebrafish. *PLoS ONE* **9**, e92217 (2014).
54. Smyth, I. *et al.* The extracellular matrix gene *Frem1* is essential for the normal adhesion of the embryonic epidermis. *Proc. Natl Acad. Sci. USA* **101**, 13560–13565 (2004).
55. Scott, D. On optimal and data-based histograms. *Biometrika* **66**, 605–610 (1979).
56. Woo, K. & Fraser, S. E. Order and coherence in the fate map of the zebrafish nervous system. *Development* **121**, 2595–2609 (1995).
57. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**, 268–269 (2011).



Extended Data Figure 1 | gDNA versus gDNA–mRNA detection of scars. **a**, Scar percentage per cell (top bar indicates clones), and pie chart of fraction of cells per clone (colours matching histograms' top bar) detected via gDNA (ScarTrace without step 1) and gDNA–mRNA detection (full protocol). **b**, Number of detected cells per clone in gDNA versus gDNA–mRNA detection and Pearson's correlation coefficient computed using the 20 different clones identified. Dot sizes are proportional to the total number of cells found taking together the two detection strategies. **c**, Fraction of cells detected per clone in gDNA (green) and gDNA–mRNA (purple) detection, in which clone '0' represents unscarred cells. **d**, Top, one-sided Fisher's exact test on a contingency table made of the number of cells detected for the given scar clone *x* for each detection strategy (a and b, respectively), and the number of cells taking together all other clones (c and d, respectively) found in gDNA detection (*n* = 147 cells in

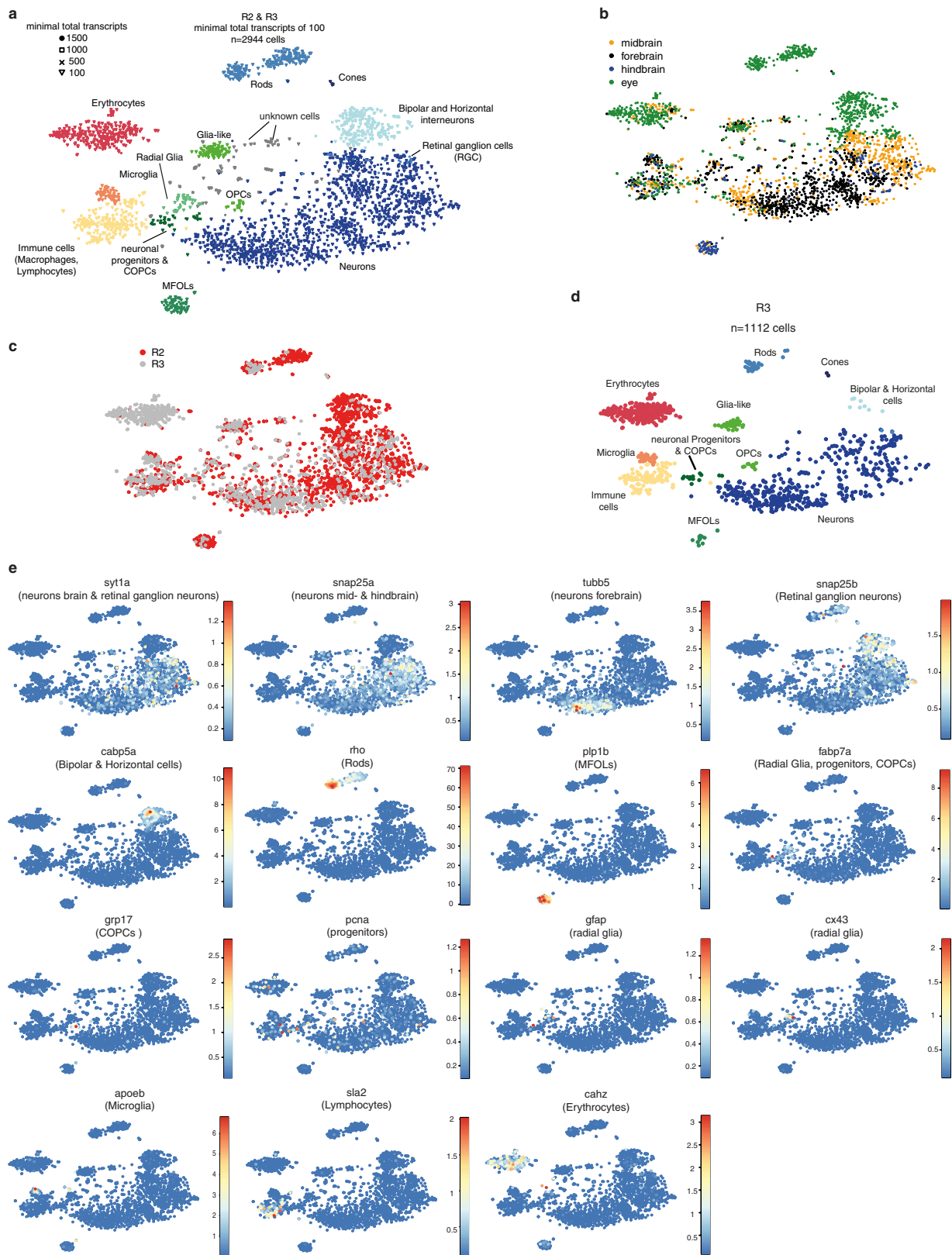
total) and gDNA–mRNA detection (*n* = 128 cells in total). Bottom, heat map shows the one-sided *P* value of each scar clone to be enriched in the gDNA or mRNA/gDNA detection protocol. No enrichment is found with *P* < 0.05, therefore results found for the two detection protocols are compatible. **e**, Normalized histograms and corresponding fit noise model function (grey line; Supplementary Information section 4) for the scar percentage detected for 1 clone found in fish P1. Scar detection efficiency is defined as the area above 5% scar content (vertical red line). Efficiency of detection of unscarred molecules is assumed to be the same as for scarred molecules. **f**, Normalized histogram of the scar detection efficiencies found after pooling all clones from all organs for all fish (in total, *n* = 371 detected clones; Extended Data Table 1). The vertical black line and the grey area indicate the mean scar detection efficiencies and s.e.m., respectively.



Extended Data Figure 2 | Transcriptome analysis of the zebrafish WKM.

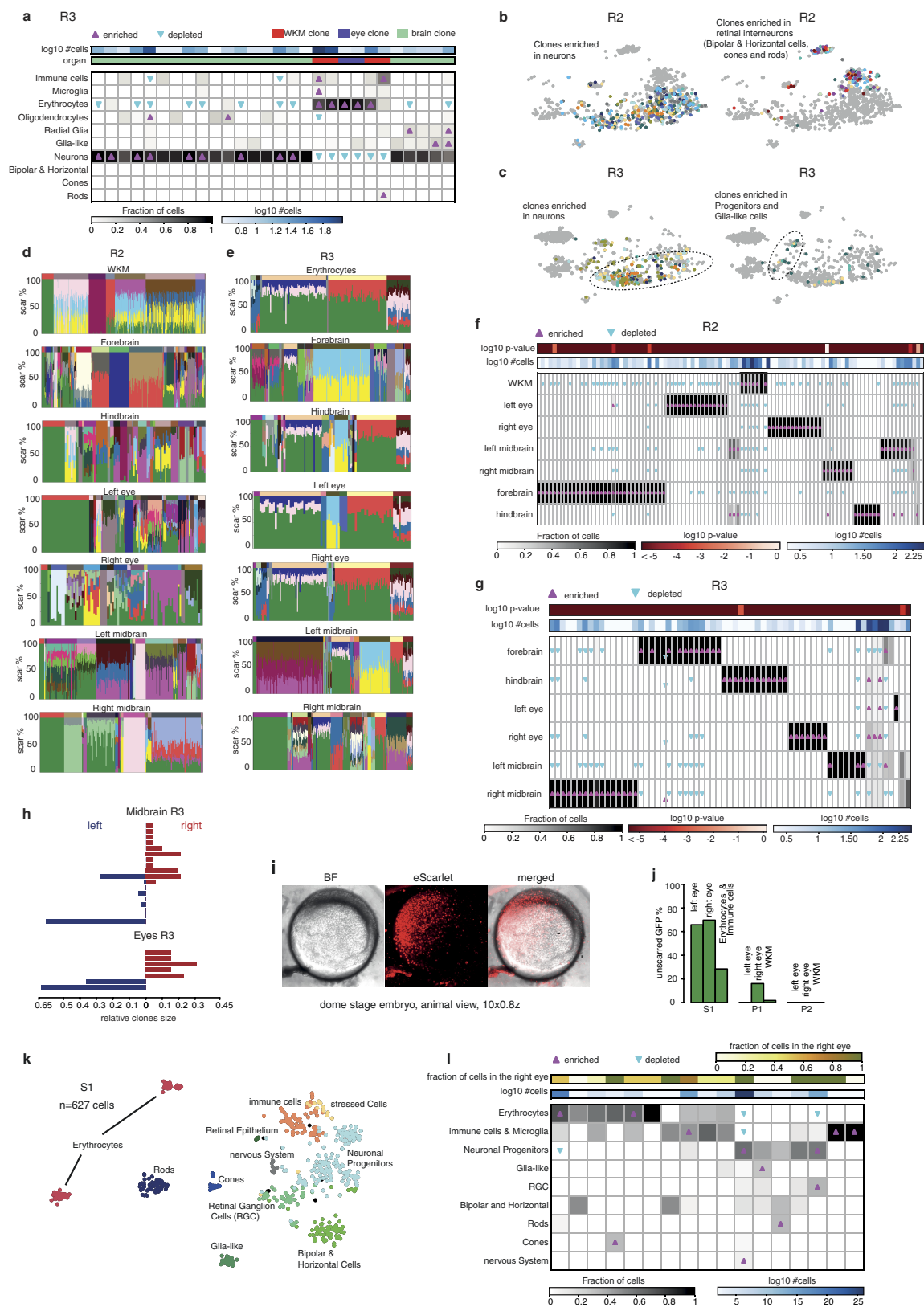
a, b, Scar percentage per cell for fish P2 (**a**) and R2 (**b**). The bar above each panel indicates clones and corresponding *P* values. **c, d**, Lineage trees for clones detected in P2 (**c**) and R2 (**d**) obtained as described in Fig. 2. **e**, *t*-SNE map of cells from fish R1 and R2. Colours and numbers indicate RaceID clusters. **f**, Heat map of the Pearson correlation between cells sorted according to RaceID clusters. Cluster numbers are indicated on the *x* and *y* axes. **g**, *t*-SNE map for the WKM of R1 and R2 coloured according

to fish of origin (R1 in pink and R2 in green). All cell types intermingle well, except erythrocytes. Even though erythrocytes appear separately on the *t*-SNE space, they belong to the same RaceID cluster. **h**, *t*-SNE maps for R1 and R2, coloured according to the number of unique transcripts per single cell for each marker^{38–41}. A full list of marker genes for each cell type is available in Supplementary Table 3. **i**, *t*-SNE map for fish R2 with cells coloured according to clone. **j**, Clonal cell fraction per cell type for fish R2.



Extended Data Figure 3 | Cell types and batch effects in the brain and eyes for fish R2 and R3. **a**, *t*-SNE map obtained with RaceID of pooled cells with a minimum of 100 total transcripts from fish R2 and R3 (isolated from the right and left eyes, right and left midbrain, forebrain and hindbrain). Different symbols indicate cells with different minimal total transcript counts. Cells are coloured according to the assigned cell type using the lowest cut off (that is, taking into account cells with at least 100 transcripts). We do not lose any cell type cluster when applying higher transcript cut offs, nor do we generate new clusters of low transcript cells

when applying lower cut offs. The fraction of cells that would be termed a different cell type with a higher cut off is very low (<1%). Low transcript cells cluster mainly around the clusters formed by high transcript cells. **b**, **c**, *t*-SNE maps as in **a**, in which cells are coloured according to organ (**b**) and fish (**c**) of origin. **d**, *t*-SNE map as in **a**, but showing only cells from fish R3, with corresponding cell types indicated. **e**, *t*-SNE maps for fish R2 and R3 coloured according to the number of unique transcripts per single cell^{42–51}. A full list of marker genes for each cell type is available in Supplementary Table 4.

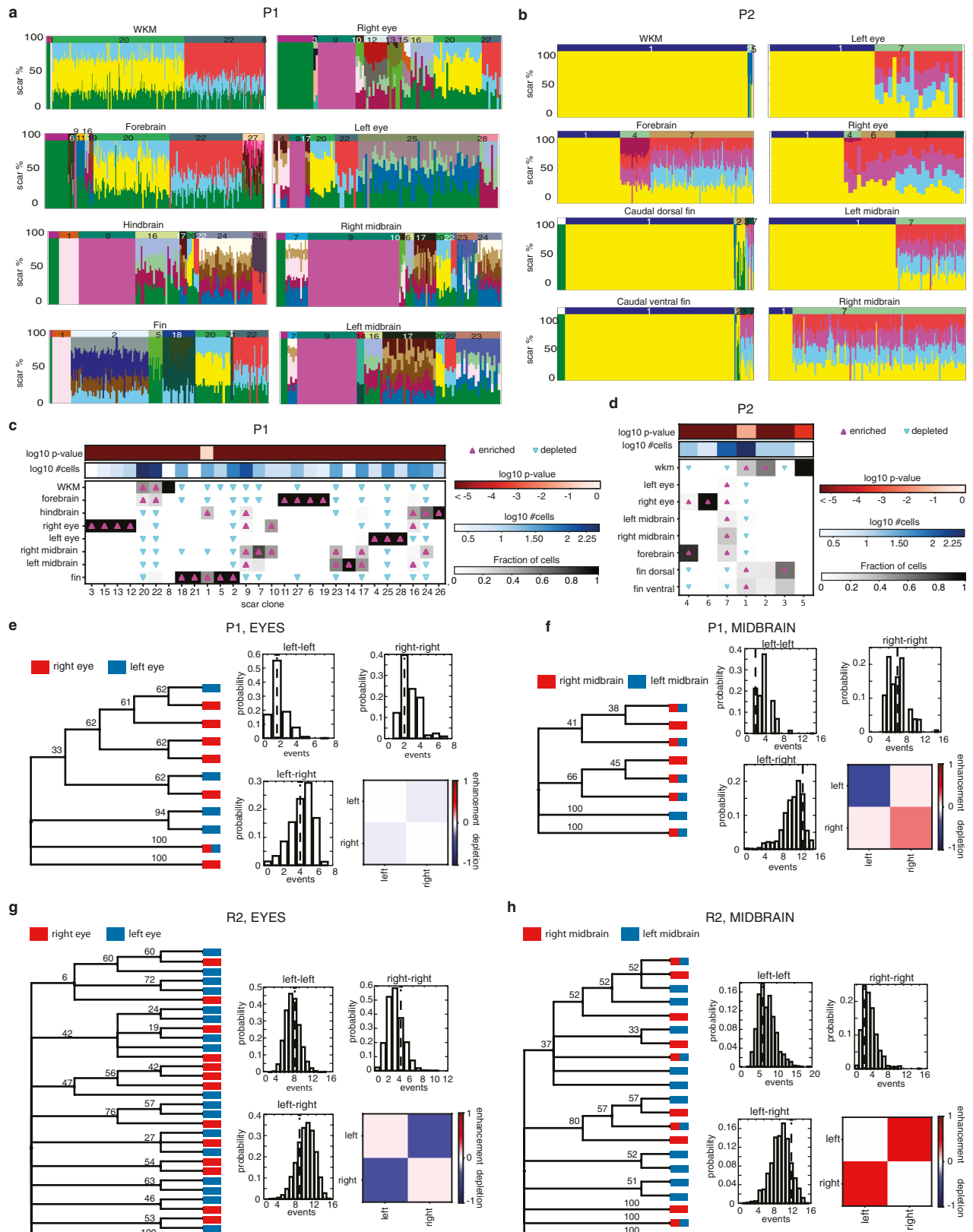


Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | ScarTrace in the zebrafish brain and eyes.

a, Heat map of the fraction of cells per clones for cell types in fish R3 (COP, OPC and MFOL clones merged as oligodendrocytes), and two-sided Fisher's exact test for enriched (magenta upwards triangle) and depleted (blue downwards triangle) clones per cell type with $P < 0.05$. The bars at the top depict organ and corresponding total number of cells. All clones have P values $< 10^{-5}$. **b, c**, t -SNE map of fish R2 and R3 cells showing different colours for enriched clones detected in glia cells, neurons or retinal interneurons. Other cells are shown in grey. **d, e**, Scar percentage per cell for clones found in the WKM, forebrain, hindbrain, left and right eyes, and left and right midbrain for R2 and R3. In all panels, each colour represents the same scar (for example, the yellow scar is the same for R2 and R3), and unscarred GFP is shown in green. **f, g**, Heat maps of the fraction of cells per clones for each organ for R2 (**f**) and R3 (**g**). Enriched (magenta upwards triangles) and depleted (blue downwards triangles)

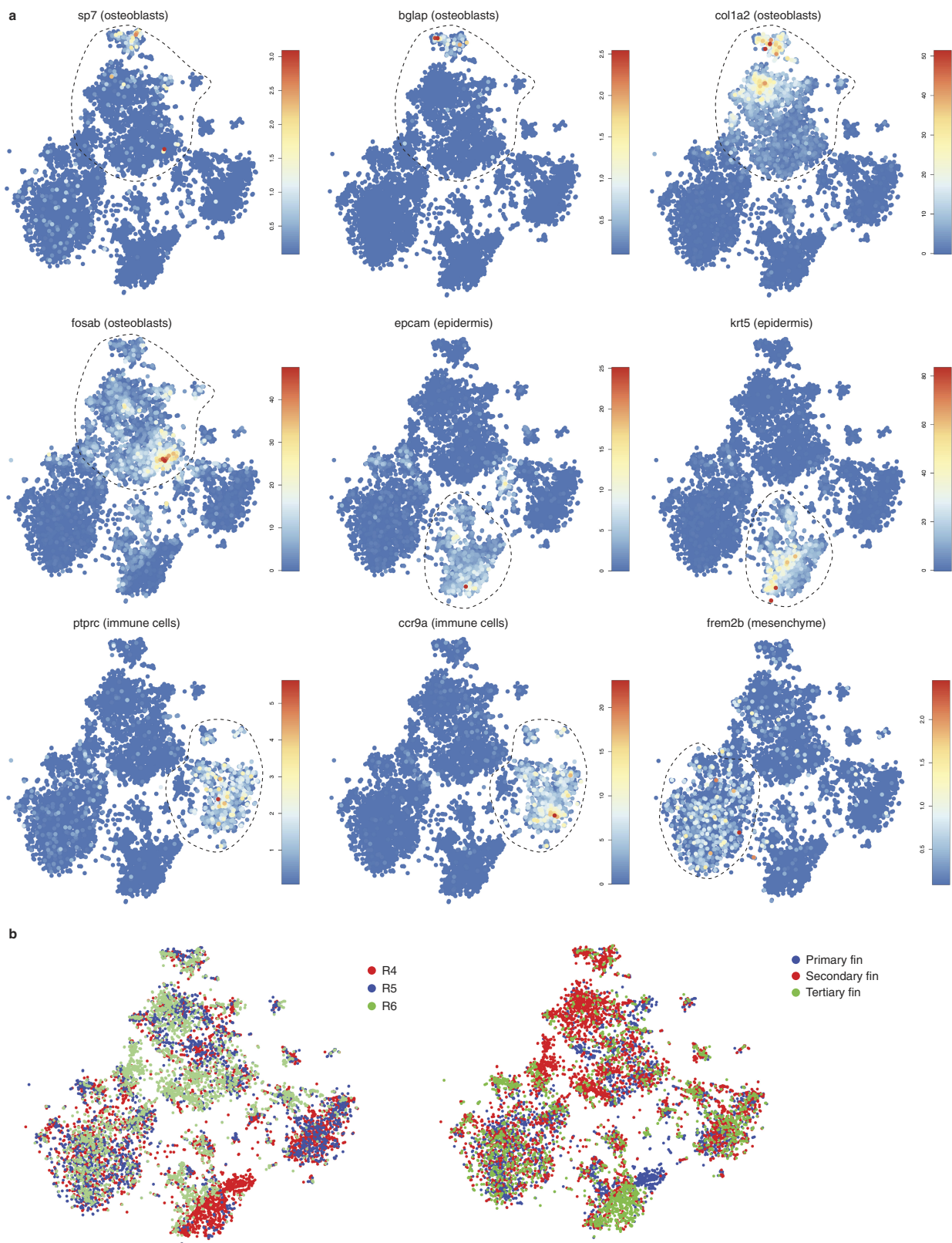
scar clones per organ are determined by a two-sided Fisher's exact test with $P < 0.05$. The bar above each panel depicts the number of cells and P value for each clone. **h**, Histograms of the relative clone frequency in the left (blue) and right (red) midbrain and eye for R3. **i**, Image of dome-stage embryo injected with Cas9-eScarlet in one cell at the two-cell stage ($n > 10$ embryos showed similar patterns). BF, bright-field. **j**, Scarring efficiency shown as the percentage of unscarred GFP in S1, P1 and P2 for the left and right eyes, and the WKM. **k**, t -SNE map of cells isolated from the left and right eyes of S1, in which cells are coloured according to their cell type. **l**, Heat map of the fraction of cells per clones for each cell type in S1. Enriched (magenta upwards triangle) and depleted (blue downwards triangle) scar clones per cell type are determined from a two-sided Fisher's exact test with $P < 0.05$. The bars at the top depict the total number of cells and the fraction of cells found in the right eye for each clone. All P values are below 10^{-5} .



Extended Data Figure 5 | See next page for caption.

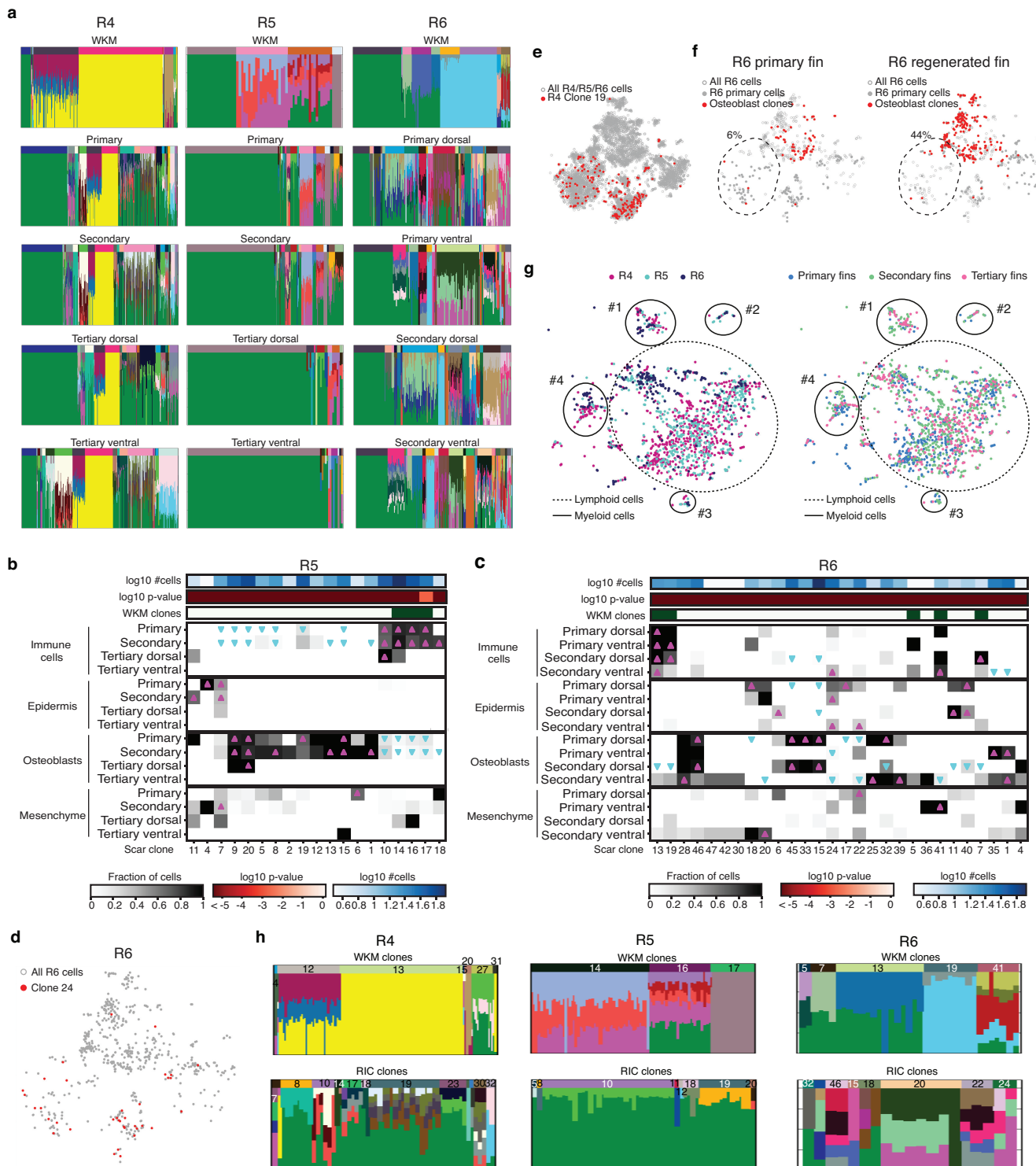
Extended Data Figure 5 | Clones for fish P1 and P2 and lineage tree of the eyes and midbrain. **a, b**, Scar percentage per cell for clones found in the WKM, forebrain, hindbrain, left and right eyes, and left and right midbrain for P1 (**a**) and P2 (**b**). Each colour represents a different scar, in which unscarred GFP is always shown in green. Colour legend per scars is different between panels (yellow scar in **a** is not yellow scar in **b**). **c, d**, Heat maps of the fraction of cells per clones for each organ for P1 (**c**) and P2 (**d**). Enriched (magenta upwards triangles) and depleted (blue downwards triangles) scar clones per organ are determined from a two-sided Fisher's exact test with $P < 0.05$. The bar above each panel depicts the number of cells and P value for each clone. **e–h**, Lineage trees obtained assuming the principle of maximum parsimony as described in Supplementary Information section 5 for clones detected in the right and the left eyes (**e, g**) and right and left midbrain (**f, h**) of fish P1 (**e, f**) and P2 (**g, h**). The root of the trees is set as an unscarred clone, with eight copies of the GFP transgene. In the tips there are the detected clones. The statistical confidence of each branch is computed as the proportion of each branch among 10,000 tree replicates constructed by bootstrapping scars

present in all clones. To assess statistically whether clones from the left or the right side co-evolve together, we randomized the clones at the tips of the tree and checked how many times, randomly, clones from the right or the left were found to be sisters with other clones from the right or the left. This allowed us to build a distribution of co-evolution (histograms in each tree) of clones for the null hypothesis and check whether the number of times we saw clones from one side together was statistically significant or not. The vertical dashed line in each histogram indicates the number of times we see clones from one side together as sisters in the reference tree. When such line is found at the right-hand (left-hand) side of the maximum, we assume that the coevolution of the clones is enhanced (depleted). In the heat maps, we indicate the degree of co-evolution of clones in the right or the left eye or midbrain, computed as the fraction of the area of the histogram at the right- or the left-hand side (that is, enhanced or depleted co-evolution, respectively) of the vertical line divided by the corresponding area of the histogram at the right- or left-hand side of maximum of the distribution.



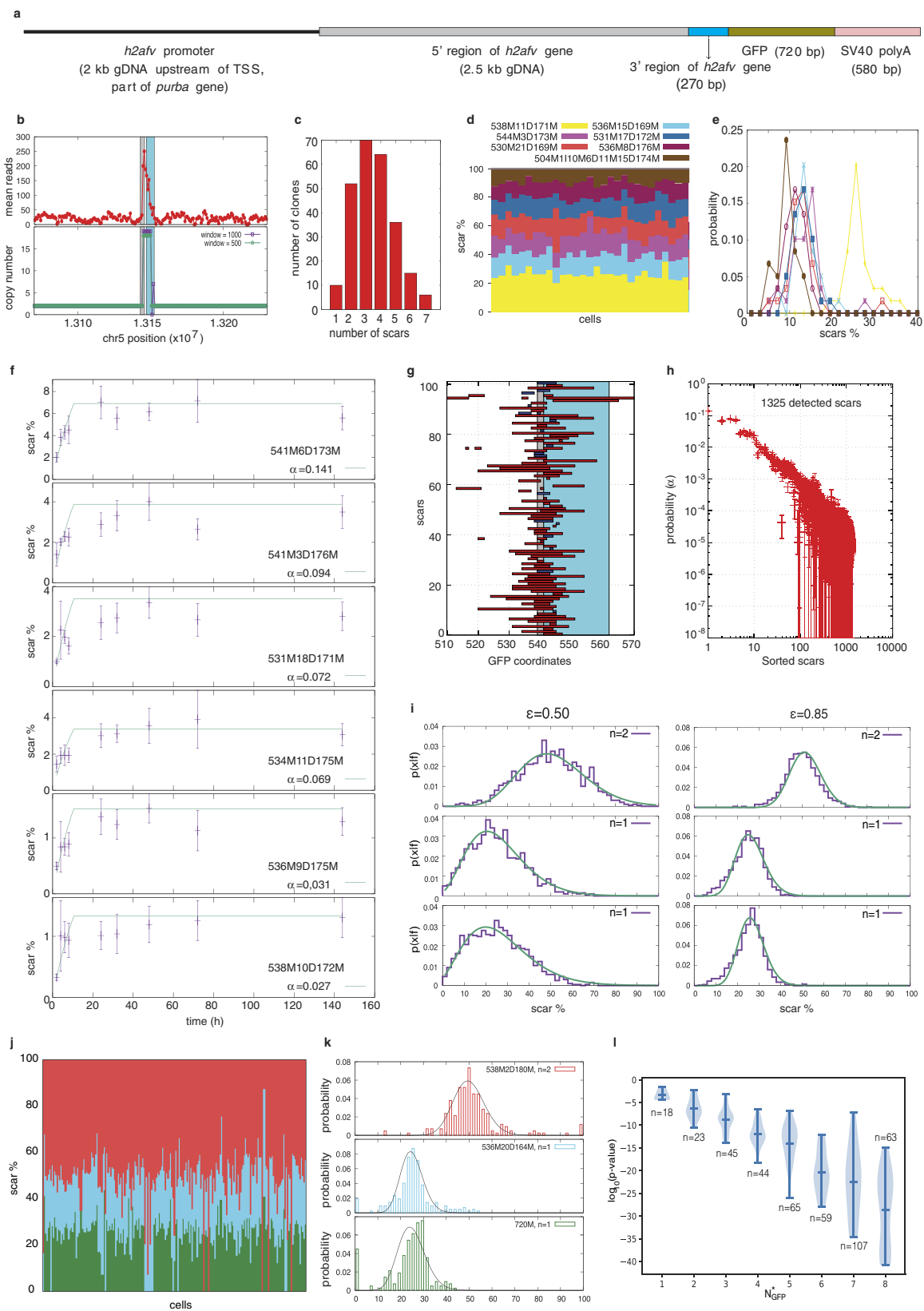
Extended Data Figure 6 | Transcriptome analysis of the zebrafish caudal fin. **a**, *t*-SNE maps obtained by pooling together cells from primary and regenerated fins from fish R4, R5 and R6. In each panel, single cells are coloured according to the number of unique transcripts observed for a given gene. The corresponding cell type is indicated in parenthesis^{52–54}.

A complete list of marker genes used for each cell type is available in Supplementary Table 5. **b**, *t*-SNE maps for the caudal fin of R4, R5 and R6 coloured based on fish (left) and fin version (right) of origin. All cells are present in all fins and fin version. No batch effects are observed.



Extended Data Figure 7 | Clonal analysis in the caudal fin. **a**, Scar percentage per cell in clones detected in fish R4 (left), R5 (middle) and R6 (right). The corresponding organ is indicated above each barplot (WKM or fin version). Spatial information (dorsal or ventral) is indicated when available. The bars at the top indicate clones. Each colour represent a scar, the same colour scheme is used for all panels. **b**, **c**, Heat maps of the fraction of cells per clones for each cell type and fin in fish R5 (**b**) and R6 (**c**). Enriched (magenta upwards triangle) and depleted (blue downwards triangle) scar clones per cell type per primary, secondary and tertiary fin obtained from two-sided Fisher's exact test with $P < 0.05$. Top bars depict clones found in the WKM of the same fish, the corresponding number of cells, and the P value for each clone. **d**, t -SNE map of R6, in which cells with clone 24 (as a representative example of clones shared between mesenchymal and epidermal cells) are highlighted in red. **e**, t -SNE map

of all cells detected in the caudal fin, in which cells from clone 19 (as a representative example of clones shared between mesenchymal and epidermal cells) in R4 are highlighted in red. **f**, t -SNE map of R6 primary (left) and regenerated (right) caudal fin cells (grey circles), in which cells from osteoblast clones are highlighted in red. Dashed lines represent mesenchymal cells (Fig. 4b, Extended Data Fig. 6). The percentages indicate the fraction of mesenchymal cells that share clones with osteoblasts. **g**, Magnified view of the t -SNE maps of R4, R5 and R6 for immune cells (dashed line on Fig. 4b). Cells are coloured based on fish (left) and fin version (right) of origin. Subpopulations of lymphoid (dashed circles) and myeloid (solid circles) are found in all fish and fin versions. **h**, Scar percentage for cells detected in the WKM (top) and RICs (bottom) for R4 (left), R5 (middle) and R6 (right) in the primary fins reveals the absence of common scars between the two. The bar above each panel indicates the different clones.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | The histone-GFP transgene and scar

characterization. **a**, Scheme of one copy of the h2afva:GFP transgene as previously described¹⁸. **b**, Copy number of the transgene. Top, average number of reads in bin sizes of 1 kb and sliding window of 200 bp obtained in whole-genome sequencing data. Bottom, copy number extracted using FREEC-11.0 with default parameters (Methods). **c**, Number of clones detected with a given number of scars, obtained by pooling all data from all fish used in this study. **d**, Scar percentage per cell in a clone in which seven different scars are detected. **e**, Probability density function (normalized histogram) of the fraction of scars detected in the clone depicted in **d** (colour code as in **d**). **f**, Average fraction of a scar per embryo computed over ten independently injected embryos (for times larger than 6 h) and over three pools of ten embryos (for times lower or equal to 6 h) detected from gDNA as a function of time for *Cas9* RNA injections, for the six most observed scars (described with CIGAR strings). Error bars denote s.e.m. Solid green lines are the fit to equation (5) in Supplementary Information section 2. **g**, Top 100 observed scars, sorted according to their probability, and corresponding position of deletions (red) and insertions

(blue) along the GFP coordinate. **h**, Scar probabilities as a function of sorted scars. Error bars denote s.e.m. from the fit (in **f**). **i**, Probabilities of measuring the percentages x_i for three scars with copy numbers 2 (top), 1 (middle) and 1 (bottom), in which the expected percentages are $f_i = 50\%$, 25% and 25% , respectively, present in the same cell with four surviving integrations of GFP. The probability has been obtained by independently simulating 1,000 times the ScarTrace protocol for $\varepsilon = 0.50$ (left) and $\varepsilon = 0.85$ (right). Solid green lines are the fit to equation (7) in Supplementary Information section 3. **j**, Scar percentage for cells from the same clone made of three scars, in which each scar is represented with a different colour. **k**, Corresponding probability density functions (normalized histogram) for the fraction of each scar per cell (colour code as in **j**). Black lines denote the best fit for each scar to equation (7) (see Supplementary Information). **l**, Violin plot showing the distribution of measured P values obtained using a Gaussian kernel density with bandwidth determined using the Scott method⁵⁵, for all clones with a given estimated N_{GFP} . Labels indicate the number of clones observed for N_{GFP} .

Extended Data Table 1 | Clone number in different organs

		WKM	Left eye	Right eye	Forebrain	Hindbrain	Left midbrain	Right midbran	Fin
RNA	R1	14	-	-	-	-	-	-	-
	R2	7	18	14	34	17	15	10	-
	R3	-	6	11	20	16	14	22	-
	R4	10	-	-	-	-	-	-	25
	R5	3	-	-	-	-	-	-	17
	R6	8	-	-	-	-	-	-	39
prot.	P1	4	5	5	2	5	6	7	4
	P2	3	1	3	2	-	1	1	2

Number of clones in different organs for different fish, for the two different Cas9 delivery strategies (RNA or protein). Unscarred clones are excluded and only clones with two or more cells are considered. Whenever clones from the WKM are known for a fish (R1, R2, R4, R5, R6, P1 or P2), they are excluded from the clones detected in the other organs.

Extended Data Table 2 | Overview of the scarred fish and organs dissected for each

fish	Cas9	age (months)	organ	Scars	cells Transcripts	Combined
P1	protein	18	WKM	286		
			forebrain	209		
			hindbrain	123		
			left eye	98		
			right eye	106		
			left midbrain	95		
			right midbrain	152		
			c. fin (1)	186		
P2	protein	18	WKM	544		
			forebrain	175		
			left eye	44		
			right eye	45		
			left midbrain	167		
			right midbrain	151		
			c. fin (1)	339		
R1	RNA	18	WKM	1738	1784	951
R2	RNA	3	WKM	1673	1397	783
			forebrain	545	563	409
			hindbrain	133	179	64
			left eye	303	403	208
			right eye	95	250	78
			left midbrain	429	302	190
			right midbrain	238	135	78
R3	RNA	3	forebrain	408	363	206
			hindbrain	152		
			left eye	82		
			right eye	204	336	171
			left midbrain	463	126	77
			right midbrain	458	287	186
R4	RNA	18	WKM	625		
			c. fin (1)	979	926	479
			c. fin (2)	893	1143	538
			dorsal c. fin (3)	504	694	329
			ventral c. fin (3)	687	582	292
R5	RNA	18	WKM	66		
			c. fin (1)	1090	757	393
			c. fin (2)	1235	1022	673
			dorsal c. fin (3)	103	348	63
			ventral c. fin (3)	96	169	24
R6	RNA	18	WKM	432		
			dorsal c.fin (1)	318	576	135
			ventral c.fin (1)	205	538	66
			dorsal c.fin (2)	378	697	195
			ventral c.fin (2)	391	714	199

Number of cells sequenced per organ that survive filtering thresholds for transcriptome libraries, scars libraries and combined per scarred fish. The total number of sorted plates is indicated in parenthesis. In the caudal fin (c. fin), primary (1), secondary (2) or tertiary (3) are also indicated in parenthesis.

Itaconate is an anti-inflammatory metabolite that activates Nrf2 via alkylation of KEAP1

Evanna L. Mills^{1,2,3,4,*}, Dylan G. Ryan^{1*}, Hiran A. Prag⁵, Dina Dikovskaya⁶, Deepthi Menon¹, Zbigniew Zaslona¹, Mark P. Jedrychowski^{2,3}, Ana S. H. Costa⁷, Maureen Higgins⁶, Emily Hams⁸, John Szpyt³, Marah C. Runtsch¹, Martin S. King⁵, Joanna F. McGouran⁹, Roman Fischer¹⁰, Benedikt M. Kessler¹⁰, Anne F. McGettrick¹, Mark M. Hughes¹, Richard G. Carroll^{1,4}, Lee M. Booty^{4,5}, Elena V. Knatko⁶, Paul J. Meakin¹¹, Michael L. J. Ashford¹¹, Louise K. Modis⁴, Gino Brunori¹², Daniel C. Sévin¹³, Padraic G. Fallon⁸, Stuart T. Caldwell¹⁴, Edmund R. S. Kunji⁵, Edward T. Chouchani^{2,3}, Christian Frezza⁷, Alben T. Dinkova-Kostova^{6,15}, Richard C. Hartley¹⁴, Michael P. Murphy^{5§} & Luke A. O'Neill^{1,4§}

The endogenous metabolite itaconate has recently emerged as a regulator of macrophage function, but its precise mechanism of action remains poorly understood^{1–3}. Here we show that itaconate is required for the activation of the anti-inflammatory transcription factor Nrf2 (also known as NFE2L2) by lipopolysaccharide in mouse and human macrophages. We find that itaconate directly modifies proteins via alkylation of cysteine residues. Itaconate alkylates cysteine residues 151, 257, 288, 273 and 297 on the protein KEAP1, enabling Nrf2 to increase the expression of downstream genes with anti-oxidant and anti-inflammatory capacities. The activation of Nrf2 is required for the anti-inflammatory action of itaconate. We describe the use of a new cell-permeable itaconate derivative, 4-octyl itaconate, which is protective against lipopolysaccharide-induced lethality *in vivo* and decreases cytokine production. We show that type I interferons boost the expression of *Irg1* (also known as *Acod1*) and itaconate production. Furthermore, we find that itaconate production limits the type I interferon response, indicating a negative feedback loop that involves interferons and itaconate. Our findings demonstrate that itaconate is a crucial anti-inflammatory metabolite that acts via Nrf2 to limit inflammation and modulate type I interferons.

Macrophages have a key role in innate immunity. They respond rapidly to pathogens and subsequently promote an anti-inflammatory phenotype to limit damage and promote tissue repair. The factors driving these changes are incompletely understood. Itaconate, a metabolite synthesized by the enzyme encoded by *Irg1*¹, is increased in lipopolysaccharide (LPS)-activated macrophages² and has been suggested to limit inflammation by inhibiting succinate dehydrogenase (SDH), a crucial pro-inflammatory regulator⁴; however, the details remain unclear.

Itaconate was the most abundant metabolite in LPS-treated human macrophages (Fig. 1a) and reached 5 mM in mouse bone marrow-derived macrophages (BMDMs) after LPS stimulation (Fig. 1b, c). Itaconate can disrupt SDH activity, but is less potent than the classic SDH inhibitor malonate (Extended Data Fig. 1), suggesting that it may exert its anti-inflammatory effects via additional mechanisms.

Itaconate contains an electrophilic α,β -unsaturated carboxylic acid that could potentially alkylate protein cysteine residues by a Michael addition to form a 2,3-dicarboxypropyl adduct. An attractive candidate

protein that undergoes cysteine alkylation is KEAP1, a central player in the anti-oxidant response (Fig. 1d). KEAP1 normally associates with and promotes the degradation of Nrf2, but alkylation of crucial KEAP1 cysteine residues allows newly synthesized Nrf2 to accumulate, migrate to the nucleus and activate a transcriptional anti-oxidant and anti-inflammatory program⁵. We therefore examined KEAP1 and Nrf2 as targets of itaconate.

The cell-permeable itaconate derivative dimethyl itaconate (DMI)³ boosted levels of Nrf2 protein, expression of downstream target genes, including *Hmox1*, and glutathione (GSH) (Extended Data Fig. 2a–d). However, the lack of a negative charge on the conjugated ester group in DMI increases its reactivity towards Michael addition, making it a far superior Nrf2 activator than itaconate akin to the potent Nrf2 activator dimethylfumarate (DMF)⁶. DMI is rapidly degraded within cells without releasing itaconate⁷, hence is unlikely to mimic endogenous itaconate. Even so, these data indicate that Nrf2 activation is anti-inflammatory⁸ (Extended Data Fig. 2e, f).

To overcome the limitations of DMI, we synthesized 4-octyl itaconate (OI), a cell-permeable itaconate derivative (Extended Data Fig. 3a). Itaconate and OI had similar thiol reactivity that was far lower than that of DMI (Extended Data Fig. 3b, c, f), making it a suitable cell-permeable itaconate surrogate. Furthermore, OI was hydrolysed to itaconate by esterases in mouse myoblast C2C12 cells (Extended Data Fig. 3d) and LPS-activated mouse macrophages (Extended Data Fig. 3e). OI boosted Nrf2 levels (Fig. 1e, compare lane 5 to lane 1) and enhanced LPS-induced Nrf2 stabilization (Fig. 1e, compare lane 6 to lane 2), and increased the expression of downstream target genes⁹, including the anti-inflammatory protein HMOX1¹⁰ (Fig. 1f, g). We used a quantitative NAD(P)H:quinone oxidoreductase-1 (NQO1) inducer bioassay^{11,12}, to assess the potency of Nrf2 activation by the CD value (concentration required to double the specific enzyme activity) for NQO1, the prototypical Nrf2 target gene. OI (CD value of 2 μ M), was more potent than the clinically used Nrf2 activator DMF (CD value of 6.5 μ M) (Fig. 1h, Extended Data Fig. 3f). OI stimulated synthesis of the key anti-oxidant GSH (Extended Data Fig. 3g–i). OI also boosted canonical activation of Nrf2 by the pro-oxidant hydrogen peroxide (H_2O_2) (Extended Data Fig. 3j, k). Importantly, the related octyl esters 4-octyl 2-methylsuccinate and octyl succinate, which are not Michael acceptors, had no effect on Nrf2 activity, confirming the requirement

¹School of Biochemistry and Immunology, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin, Ireland. ²Department of Cancer Biology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. ³Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴GlaxoSmithKline, Gunnelswood Road, Stevenage, Hertfordshire, UK. ⁵MRC Mitochondrial Biology Unit, University of Cambridge, Cambridge CB2 0XY, UK. ⁶Jacqui Wood Cancer Centre, Division of Cancer Research, School of Medicine, University of Dundee, Dundee DD1 9SY, UK. ⁷MRC Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre, Box 197, Cambridge Biomedical Campus, Cambridge CB2 0XZ, UK. ⁸School of Medicine, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin, Ireland. ⁹School of Chemistry, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin, Ireland. ¹⁰Nuffield Department of Medicine, Target Discovery Institute, University of Oxford, Oxford OX3 7FZ, UK. ¹¹Division of Molecular and Clinical Medicine, School of Medicine, University of Dundee, Dundee DD1 9SY, UK. ¹²GlaxoSmithKline, Park Road, Ware, Hertfordshire, UK. ¹³Cellzome, GlaxoSmithKline R&D, Heidelberg, Germany. ¹⁴WestCHEM School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK. ¹⁵Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

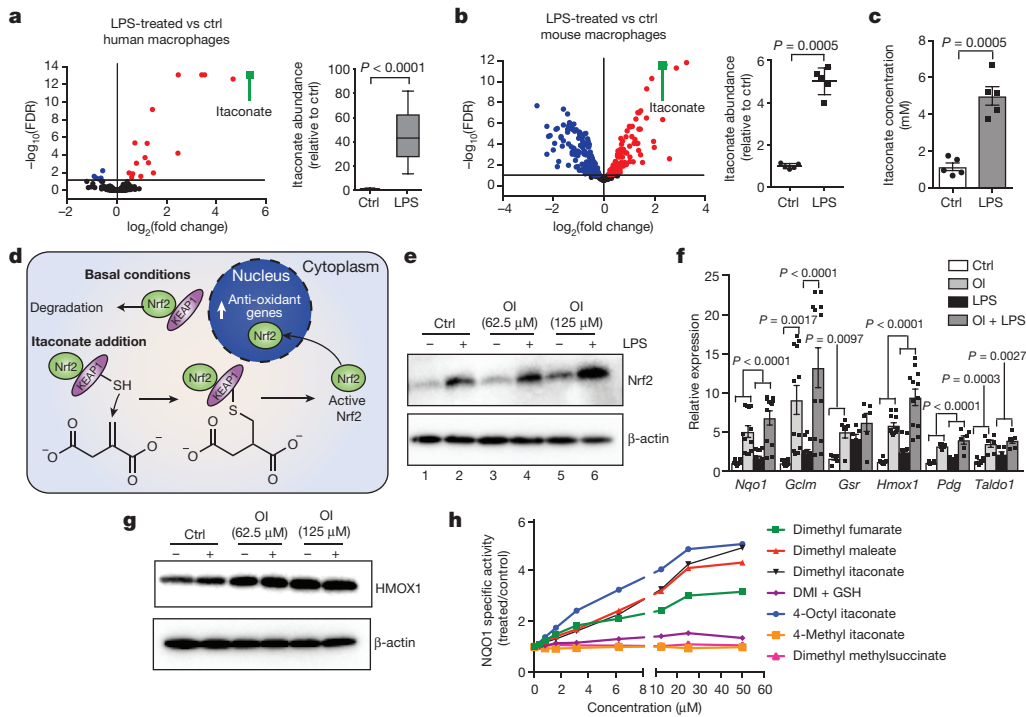


Figure 1 | Itaconate activates Nrf2. **a–c**, Metabolite levels and itaconate abundance in control (ctrl) versus LPS-induced (**a**, $n = 12$, 4 h; **b**, **c**, $n = 5$, 24 h) human (**a**) and mouse (**b**, **c**) macrophages. Red and blue dots represent metabolites significantly up- and downregulated by LPS, respectively. FDR, false discovery rate. **d**, Reactivity of itaconate with KEAP1 thiol group. **e**, **g**, LPS-induced Nrf2 (**e**, 24 h) and HMOX1 (**g**, 6 h) after treatment with OI as indicated. **f**, Nrf2 target gene expression in mouse macrophages with or without LPS (6 h) and OI (*Nqo1*, *Gclm*, *Hmox1*, $n = 12$; *Gsr*, *Pdg*, *Taldo1*, $n = 6$). **h**, NQO1 activity in mouse Hepa1c1c7 cells treated as indicated (48 h, $n = 8$). Data are mean \pm s.e.m. P values calculated using one-way or two-way analysis of variance (ANOVA) for multiple comparisons or two-tailed Student's t -test for paired comparisons. Blots are representative of three independent experiments. In the box plots, line shows mean. For gel source data, see Supplementary Fig. 1.

for the itaconate moiety (Extended Data Fig. 3l). Dimethyl malonate, a potent SDH inhibitor⁴, did not activate Nrf2 (Extended Data Fig. 3m), confirming that Nrf2 activation by OI is independent of SDH inhibition.

Itaconate is generated by IRG1 in the mitochondrial matrix and must cross the mitochondrial inner membrane to act on Nrf2 in the cytosol.

Itaconate is structurally similar to malate, which is transported across the mitochondrial inner membrane by the dicarboxylate, citrate and oxoglutarate carriers. All three carriers transported itaconate, whereas other tested carriers could not (Fig. 2a and Extended Data Fig. 4), suggesting that LPS-induced itaconate is generated in the mitochondrial matrix and is then exported to the cytosol to activate Nrf2.

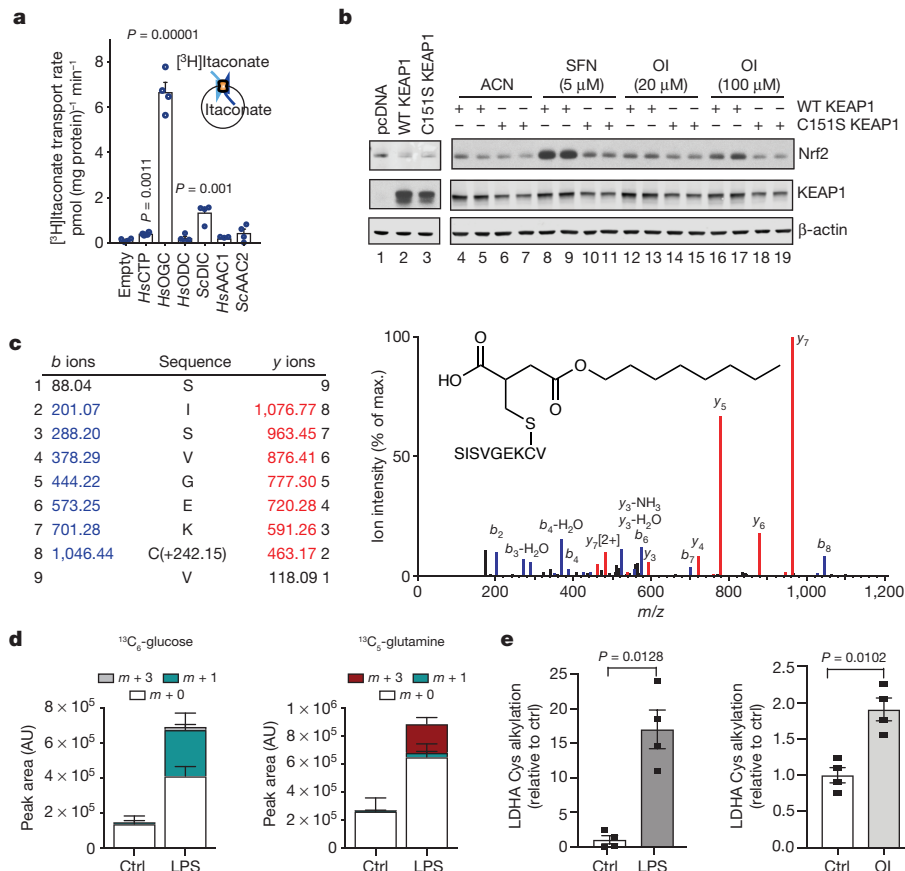


Figure 2 | Itaconate alkylates cysteines.

a, Itaconate transport by the indicated carriers ($n = 4$). *HsAAC1*, *Homo sapiens* ADP/ATP carrier; *HsCTP*, *H. sapiens* citrate carrier; *HsODC*, *H. sapiens* oxodicarboxylate carrier; *HsOGC*, *H. sapiens* 2-oxoglutarate carrier; *ScAAC2*, *Saccharomyces cerevisiae* ADP/ATP carrier; *ScDIC*, *S. cerevisiae* dicarboxylate carrier. **b**, Nrf2 and KEAP1 protein after co-transfection with Nrf2-V5, and the wild-type (WT) or Cys151Ser mutant KEAP1. **c**, Tandem mass spectrometry spectrum of Cys151-containing KEAP1 peptide after OI treatment. **d**, Metabolite ($^{13}\text{C}_6$ -glucose (left), $^{13}\text{C}_5$ -glutamine (right)) tracing to itaconate-cysteine adduct with or without LPS (24 h, $n = 5$). AU, arbitrary units. **e**, LDHA Cys84 alkylation plus LPS (24 h) or OI (250 μM , 4 h) ($n = 4$). Data are mean \pm s.e.m. (in **d**, **e**) or s.d. (in **a**). P values calculated using one-way ANOVA for multiple comparisons or two-tailed Student's t -test for paired comparisons. Blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.

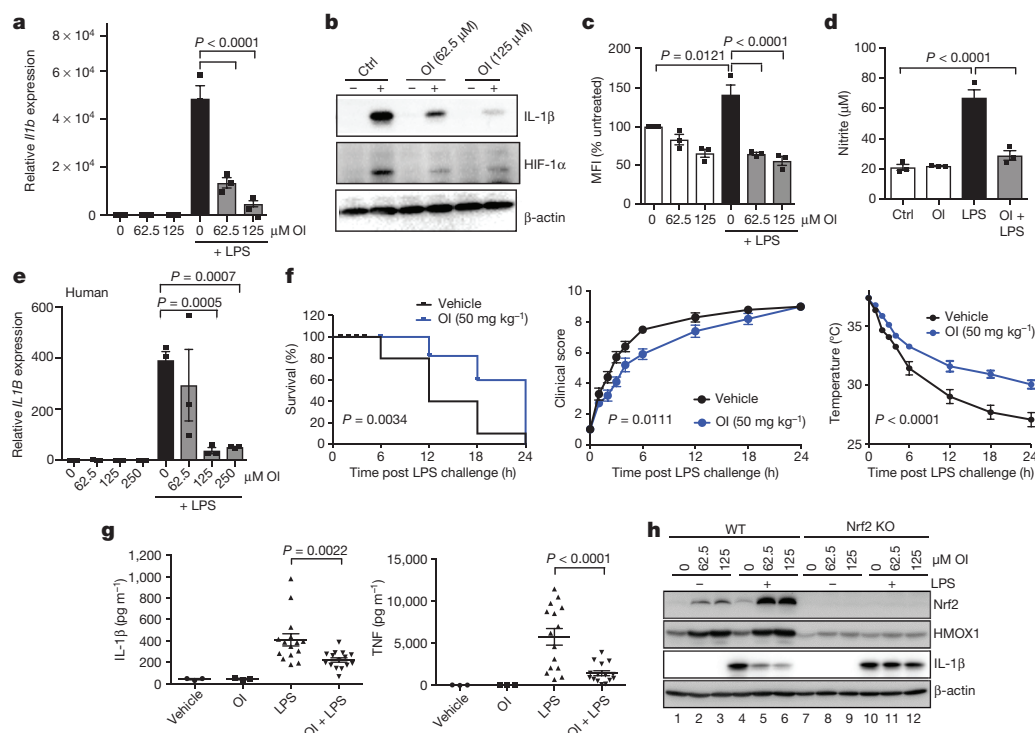


Figure 3 | OI limits IL-1 β in an Nrf2-dependent manner and protects against LPS lethality. **a–d**, LPS (24 h) induced *Il1b* mRNA (**a**, $n = 3$), IL-1 β and HIF-1 α protein (**b**), ROS production (**c**, $n = 3$; measured as the percentage change in mean fluorescent intensity (MFI) relative to untreated control) and nitrite production (**d**, $n = 3$) \pm OI. **e**, *IL1B* mRNA in human PBMCs treated as in **a–d** ($n = 3$). **f**, Survival (left), clinical score (middle) and body temperature (right) measurements in mice ($n = 10$) injected intraperitoneally with OI (50 mg kg⁻¹, 2 h) and LPS (15 mg kg⁻¹).

g, Serum IL-1 β and TNF levels from mice injected intraperitoneally with OI (50 mg kg⁻¹, 2 h) and/or LPS (2.5 mg kg⁻¹, 2 h, $n = 3$ vehicle, OI; $n = 15$ LPS, OI plus LPS). **h**, Nrf2, HMOX1 and IL-1 β protein in wild-type and Nrf2 knockout (KO) mouse BMDMs treated with LPS (6 h) and OI as indicated. Data are mean \pm s.e.m. P values calculated using one-way ANOVA. Blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.

Our hypothesis is that itaconate activates Nrf2 by alkylation of KEAP1 cysteine residue(s)^{13–15}, similar to the modification of cysteines by fumarate (Extended Data Fig. 5a). Cysteine 151 (Cys151) is a principal sensor on KEAP1 for sulforaphane¹⁶ and DMF¹⁷. OI stabilized V5-tagged Nrf2 (Nrf2-V5) in COS1 cells co-expressing wild-type KEAP1 but not a Cys151Ser mutant, similarly to sulforaphane (Fig. 2b, compare lanes 16 and 17 to lanes 18 and 19). To analyse KEAP1 alkylation directly, we overexpressed Myc-DDK-tagged KEAP1 in human HEK293T cells and treated the cells with OI. Tandem mass spectrometry of immunoprecipitated KEAP1 revealed that for the KEAP1 peptide (144–152), which contains Cys 151, OI treatment increased its mass by 242.15 Da, consistent with alkylation by OI (Fig. 2c). OI also modified other known KEAP1 regulatory cysteine residues (Cys257, Cys288 and Cys273) (Extended Data Fig. 5b–d, Extended Data Table 1a). Furthermore, itaconate-cysteine adducts, derived in part from glucose and glutamine (Fig. 2d and Extended Data Fig. 6), were detected in LPS-treated macrophages. These data suggest that itaconate activates Nrf2 by alkylating KEAP1 cysteine residues. We further explored cysteine alkylation induced by itaconate using an untargeted mass spectrometry approach in macrophages treated with OI, or with LPS, which increases itaconate levels. We identified several proteins that contain alkylated cysteine residues (Extended Data Table 1b, c). Notably LDHA, which has a crucial role in the regulation of glycolysis, was alkylated in OI- and LPS-treated macrophages (Fig. 2e and Extended Data Fig. 5e, f). This modification, here defined as 2,3-dicarboxypropylation, generates a stable thioether. As there are no known pathways for the removal of such post-translational modifications, modified proteins are probably degraded, suggesting that this modification will have profound effects on macrophage function.

We next assessed whether itaconate activation of Nrf2 could be anti-inflammatory. OI, used at concentrations that did not affect

cellular viability, decreased LPS-induced *Il1b* mRNA, pro-IL-1 β , HIF-1 α and IL-10 protein levels, and decreased the extracellular acidification rate, yet had no effect on NF- κ B activity or TNF (also known as TNF α) levels (Fig. 3a, b and Extended Data Fig. 7a–f). OI also decreased *Il1b* mRNA in BMDMs treated with the TLR2 and TLR3 ligands, Pam3CSK and polyinosinic:polycytidylic acid (poly(I:C)), respectively (Extended Data Fig. 7g). Levels of LPS-induced reactive oxygen species (ROS), nitrite and inducible nitric oxide synthase (iNOS) were limited by OI (Fig. 3c, d and Extended Data Fig. 7h, i). These effects are likely to be a consequence of ROS detoxification after Nrf2 induction by OI. IL-1 β and TNF were decreased by OI in human peripheral blood mononuclear cell (PBMCs) (Fig. 3e, Extended Data Fig. 7j). OI also counteracted the pro-inflammatory response to LPS *in vivo*. OI, which activated Nrf2 (Extended Data Fig. 7k), prolonged survival, decreased clinical score and improved body temperature regulation, and decreased IL-1 β and TNF levels but not IL-10 in an LPS model of sepsis (Fig. 3f, g and Extended Data Fig. 7l).

OI induction of HMOX1 was blocked in Nrf2-deficient macrophages (Fig. 3h (compare lanes 2 and 3 to lanes 8 and 9) and Extended Data Fig. 8a, d) or when Nrf2 was silenced (Extended Data Fig. 8a, d (compare lanes 7 and 8 to lanes 11 and 12)). Without Nrf2, the decrease in LPS-induced IL-1 β with OI was significantly impaired (Fig. 3h (compare lane 6 to lane 12), Extended Data Fig. 8b–f (compare lanes 6 and 8 to 10 and 12 in c, d)). Furthermore, two Nrf2 activators, diethyl maleate and 15-deoxy- $\Delta^{12,14}$ -prostaglandin J₂ decreased LPS-induced IL-1 β , IL-10, nitric oxide synthase (NOS2) and nitrite (Extended Data Fig. 8g–k). Thus, itaconate activates an anti-inflammatory program through Nrf2.

We next investigated how switching from a pro- to an anti-inflammatory state might affect itaconate production from aconitate by IRG1.

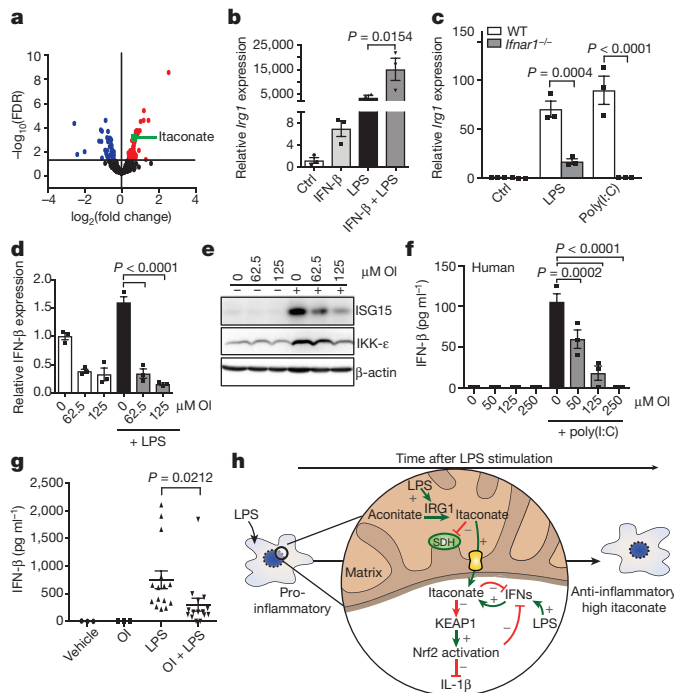


Figure 4 | A feedback loop exists between itaconate and IFN-β.

a, Metabolite levels in control versus IFN-β-treated (1,000 U ml⁻¹; 27 h; *n* = 5) mouse macrophages. **b**, LPS-induced (24 h) *Irg1* expression ± IFN-β (1,000 U ml⁻¹; *n* = 3). **c**, *Irg1* expression in wild-type and IFN receptor-deficient (*Ifnar1*^{-/-}) BMDMs plus LPS or poly(I:C) (40 μg ml⁻¹) for 24 h (*n* = 3). **d**, IFN-β (*n* = 3) expression plus LPS (24 h) and OI as indicated. **e**, ISG15 and IKK-ε expression after treatment with LPS (24 h) and OI. **f**, IFN-β protein expression in PBMCs treated with poly(I:C) (20 μg ml⁻¹; 24 h) and OI (*n* = 3) as indicated. **g**, Serum IFN-β levels from mice injected intraperitoneally with OI (50 mg kg⁻¹, 2 h) with or without LPS (2.5 mg kg⁻¹, 2 h) (*n* = 3 vehicle, OI; *n* = 15 LPS, OI and LPS). **h**, The anti-inflammatory role of itaconate. Data are mean ± s.e.m. *P* values calculated using one-way ANOVA. Blots are representative of three independent experiments. Data in **f** are representative from one of two human donors. For gel source data, see Supplementary Fig. 1.

By modelling gene networks that control *Irg1* expression, the IFN response factor IRF1 was identified as a regulator¹⁸. We show here that itaconate levels are increased after IFN-β treatment (Fig. 4a), in agreement with others¹⁹. Levels of citrate and aconitate, the substrate for *Irg1*, were reduced by IFN-β as was the downstream metabolite α-ketoglutarate (Extended Data Fig. 9a). These data are consistent with an increase in aconitate conversion to itaconate rather than α-ketoglutarate. IFN-β enhanced basal and LPS-induced *Irg1* expression (Fig. 4b). LPS- and poly(I:C)-induced *Irg1* expression in BMDMs lacking type I IFN receptor was decreased (Fig. 4c), indicating that autocrine IFN facilitates *IRG1* induction. OI limited the IFN response, decreasing the expression of IFN-β, IKK-ε, ISG20 and ISG15 protein, IFN-β production in poly(I:C)-treated PBMCs and LPS-induced IFN-β production *in vivo* (Fig. 4d–g and Extended Data Fig. 9b, c). IFN-β enhanced both the mRNA and protein expression of IL-10, with or without the addition LPS (Extended Data Fig. 9d), suggesting that the decrease in IL-10 after OI treatment is due to reduced type I IFN production²⁰. Nrf2 knockout or knockdown attenuated the reduction of ISG20 expression by OI, whereas the Nrf2 activators diethyl maleate and 15-deoxy-Δ^{12,14}-prostaglandin J2 reduced ISG20 expression (Extended Data Fig. 9e–g). This agrees with increased expression of IRF3-regulated genes in LPS-treated Nrf2-deficient mice²¹.

These data suggest the operation of a negative-feedback loop: itaconate is generated in response to LPS, in part through type I IFNs, and promotes an anti-inflammatory program by Nrf2 activation (Fig. 4h), as well as SDH inhibition^{3,22}. This limits further inflammatory

gene expression and its own production by downregulating the IFN response. This helps to explain why Nrf2-deficient mice are more sensitive to septic shock²¹, even though under certain circumstances these mice are protected from inflammation²³. Our identification of itaconate as an inflammatory regulator, that directly modifies proteins through a newly identified post-translational modification, unveils therapeutic opportunities to use itaconate or OI to treat inflammatory diseases²⁴. Furthermore, an intriguing link was recently made²⁵ from itaconate to vitamin B₁₂, and this warrants further investigation in the context of inflammation and immunity. Further understanding the role of itaconate as an anti-inflammatory metabolite and regulator of type I IFNs is likely to yield new insights into the pathogenesis of inflammatory diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 July 2017; accepted 9 February 2018.

Published online 28 March 2018.

- Michelucci, A. *et al.* Immune-responsive gene 1 protein links metabolism to immunity by catalyzing itaconic acid production. *Proc. Natl Acad. Sci. USA* **110**, 7820–7825 (2013).
- Strelko, C. L. *et al.* Itaconic acid is a mammalian metabolite induced during macrophage activation. *J. Am. Chem. Soc.* **133**, 16386–16389 (2011).
- Lamproulou, V. *et al.* Itaconate links inhibition of succinate dehydrogenase with macrophage metabolic remodeling and regulation of inflammation. *Cell Metab.* **24**, 158–166 (2016).
- Mills, E. L. *et al.* Succinate dehydrogenase supports metabolic repurposing of mitochondria to drive inflammatory macrophages. *Cell* **167**, 457–470 (2016).
- Hayes, J. D. & Dinkova-Kostova, A. T. The Nrf2 regulatory network provides an interface between redox and intermediary metabolism. *Trends Biochem. Sci.* **39**, 199–218 (2014).
- Brennan, M. S. *et al.* Dimethyl fumarate and monoethyl fumarate exhibit differential effects on KEAP1, NRF2 activation, and glutathione depletion *in vitro*. *PLoS One* **10**, e0120254 (2015).
- ElAzzouny, M. *et al.* Dimethyl itaconate is not metabolized into itaconate intracellularly. *J. Biol. Chem.* **292**, 4766–4769 (2017).
- Kobayashi, E. H. *et al.* Nrf2 suppresses macrophage inflammatory response by blocking proinflammatory cytokine transcription. *Nat. Commun.* **7**, 11624 (2016).
- Lee, J. M., Calkins, M. J., Chan, K., Kan, Y. W. & Johnson, J. A. Identification of the NF-E2-related factor-2-dependent genes conferring protection against oxidative stress in primary cortical astrocytes using oligonucleotide microarray analysis. *J. Biol. Chem.* **278**, 12029–12038 (2003).
- Piantadosi, C. A. *et al.* Heme oxygenase-1 couples activation of mitochondrial biogenesis to anti-inflammatory cytokine expression. *J. Biol. Chem.* **286**, 16374–16385 (2011).
- Prochaska, H. J. & Santamaria, A. B. Direct measurement of NAD(P)H:quinone reductase from cells cultured in microtiter wells: a screening assay for anticarcinogenic enzyme inducers. *Anal. Biochem.* **169**, 328–336 (1988).
- Fahey, J. W., Dinkova-Kostova, A. T., Stephenson, K. K. & Talalay, P. The “Prochaska” microtiter plate bioassay for inducers of NQO1. *Methods Enzymol.* **382**, 243–258 (2004).
- Dinkova-Kostova, A. T. *et al.* Direct evidence that sulfhydryl groups of Keap1 are the sensors regulating induction of phase 2 enzymes that protect against carcinogens and oxidants. *Proc. Natl Acad. Sci. USA* **99**, 11908–11913 (2002).
- McMahon, M., Lamont, D. J., Beattie, K. A. & Hayes, J. D. Keap1 perceives stress via three sensors for the endogenous signaling molecules nitric oxide, zinc, and alkenals. *Proc. Natl Acad. Sci. USA* **107**, 18838–18843 (2010).
- Dinkova-Kostova, A. T., Kostov, R. V. & Canning, P. Keap1, the cysteine-based mammalian intracellular sensor for electrophiles and oxidants. *Arch. Biochem. Biophys.* **617**, 84–93 (2017).
- Zhang, D. D. & Hannink, M. Distinct cysteine residues in Keap1 are required for Keap1-dependent ubiquitination of Nrf2 and for stabilization of Nrf2 by chemopreventive agents and oxidative stress. *Mol. Cell. Biol.* **23**, 8137–8151 (2003).
- Linker, R. A. *et al.* Fumaric acid esters exert neuroprotective effects in neuroinflammation via activation of the Nrf2 antioxidant pathway. *Brain* **134**, 678–692 (2011).
- Tallam, A. *et al.* Gene regulatory network inference of immunoresponsive gene 1 (*IRG1*) identifies interferon regulatory factor 1 (*IRF1*) as its transcriptional regulator in mammalian macrophages. *PLoS One* **11**, e0149050 (2016).
- Naujoks, J. *et al.* IFNs modify the proteome of legionella-containing vacuoles and restrict infection via *IRG1*-derived itaconic acid. *PLoS Pathog.* **12**, e1005408 (2016).
- Guarda, G. *et al.* Type I interferon inhibits interleukin-1 production and inflammasome activation. *Immunity* **34**, 213–223 (2011).

21. Thimmulappa, R. K. *et al.* Nrf2 is a critical regulator of the innate immune response and survival during experimental sepsis. *J. Clin. Invest.* **116**, 984–995 (2006).
22. Cordes, T. *et al.* Immunoresponsive gene 1 and itaconate inhibit succinate dehydrogenase to modulate intracellular succinate levels. *J. Biol. Chem.* **291**, 14274–14284 (2016).
23. Freigang, S. *et al.* Nrf2 is essential for cholesterol crystal-induced inflammasome activation and exacerbation of atherosclerosis. *Eur. J. Immunol.* **41**, 2040–2051 (2011).
24. Dinarello, C. A. Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. *Blood* **117**, 3720–3732 (2011).
25. Shen, H. *et al.* The human knockout gene CLYBL connects itaconate to vitamin B₁₂. *Cell* **171**, 771–782 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. McMahon and J. D. Hayes for plasmids, and Cancer Research UK (C20953/A18644) and the BBSRC (BB/L01923X/1) for financial support for ATDK. This work was supported by a Wellcome Trust Investigator award to R.C.H. (110158/Z/15/Z), a grant to M.P.M. from the Medical Research Council UK (MC_U105663142), a Wellcome Trust Investigator award to MPM (110159/Z/15/Z), and a grant to E.R.S.K. and M.S.K. from the Medical Research Council UK (MC_U105663139). B.M.K. and R.F. are supported by the Kennedy Trust Fund. We acknowledge Metabolon for their assistance with the metabolic work and analysis. The O'Neill laboratory acknowledges the following grant support: European Research Council (ECFP7-ERC-MICROINNATE), Science Foundation Ireland Investigator Award (SFI 12/IA/1531), GlaxoSmithKline Visiting Scientist Programme and The

Wellcome Trust (oneill-wellcometrust-metabolic, grant number 205455). E.T.C. is supported by the Claudia Adams Barr Program.

Author Contributions E.L.M. and D.G.R. designed and performed experiments and analysed the data. E.L.M. wrote the manuscript with assistance from all other authors. D.M., M.M.H., M.C.R. and A.F.M. performed *in vitro* experiments using OI. R.G.C., D.C.S., A.S.H.C. and C.F. assisted with the metabolomics analysis. Z.Z., P.G.F. and E.H. assisted with the *in vivo* mouse LPS trials. S.T.C. and R.C.H. were responsible for the design and synthesis of octyl esters. H.A.P., E.R.S.K., M.S.K. and L.M.B. assessed the effect of OI and itaconate on mitochondrial parameters and itaconate transport. D.D., M.H. and A.T.D.-K. performed the NQO1 assay and KEAP1 wild-type and Cys151Ser mutant experiments. J.F.M., R.F., B.M.K., E.T.C., M.P.J. and J.S. assisted with mass spectrometry experiments. L.K.M. and G.B. provided guidance and advice. E.V.K., P.J.M. and M.L.J.A. assisted with experiments in Nrf2-deficient mice. L.A.O'N. conceived ideas and oversaw the research programme. M.P.M. provided advice, reagents and oversaw a portion of the work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to L.A.O'N. (laoneill@tcd.ie).

Reviewer Information *Nature* thanks N. S. Chandel, R. Rossignol, S. Werner and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Isolation of human PBMCs. Human PBMCs were isolated from human blood using Lymphoprep (Axis-Shield). Whole blood (30 ml) was layered on 20 ml lymphoprep and spun for 20 min at 2,000 r.p.m. with no brake on. The PBMCs were isolated from the middle layer. PBMCs were maintained in RPMI supplemented with 10% (v/v) FCS, 2 mM L-glutamine, and 1% penicillin/streptomycin solution.

Generation of human macrophages. Blood was layered on Histopaque and centrifuged at 800g for 20 min, acceleration 9, and deceleration at 4. The PBMC layer was isolated and the macrophages were sorted using magnetic-activated cell sorting (MACS) CD14 beads. Cells were plated at 0.5×10^6 cells ml⁻¹ in media containing M-CSF (100 ng ml⁻¹) and maintained at 37°C, 5% CO₂ for 5 days, to allow differentiation into macrophages. For further details, see Supplementary Methods.

Generation and treatment of BMDMs. Mice were euthanized in a CO₂ chamber and death was confirmed by cervical dislocation. Bone marrow cells were extracted from the leg bones and differentiated in DMEM (containing 10% fetal calf serum, 1% penicillin/streptomycin and 20% L929 supernatant) for 6 days, at which time they were counted and replated for experiments. Unless stated, 5×10^6 BMDMs per millilitre were used in *in vitro* experiments. Unless stated, the LPS concentration used was 100 ng ml⁻¹, the DMI and OI concentration was 125 μM, and in experiments where pre-treatments occurred before LPS stimulation this was for 3 h.

Synthesis of itaconate compounds. For details on synthesis and characterization of chemical compounds, see Supplementary Methods.

Metabolomic analysis with Metabolon. Macrophages were plated at 2×10^6 per well in 6-well plates and treated as required. BMDMs $n = 5$, human macrophages $n = 12$. Analysis was performed by Metabolon. For further details, see Supplementary Methods.

Metabolite measurements for absolute succinate and itaconate quantification and metabolite tracing. Cells were treated as desired. For tracing studies, immediately before LPS stimulation, the media was removed and replaced with DMEM media (1 ml) containing U-¹³C-glucose (4.5 g l⁻¹) or U-¹³C-glutamine (584 mg ml⁻¹) deplete of ¹²C-glucose or ¹²C-glutamine. Samples were extracted in methanol/acetonitrile/water, 50:30:20 (v/v/v) (1 ml per 1×10^6 cells) and agitated for 15 min at 4°C in a Thermomixer and then incubated at -20°C for 1 h. Samples were centrifuged at maximum speed for 10 min at 4°C. The supernatant was transferred into a new tube and centrifuged again at maximum speed for 10 min at 4°C. The supernatant was transferred autosampler vials. Liquid chromatograph-mass spectrometry (LC-MS) analysis was performed using a Q Exactive mass spectrometer coupled to a Dionex U3000 UHPLC system (Thermo). For further details, see Supplementary Methods.

Western blotting. Protein samples from cultured cells were prepared by direct lysis of cells in 5× Laemmli sample buffer, followed by heating at 95°C for 5 min. For spleen samples, 30 mg of spleen was homogenized in RIPA buffer using the Qiagen TissueLyserII system. The resulting homogenate was centrifuged at 14,000 r.p.m. for 10 min at 4°C, and supernatants were used for SDS-PAGE. Protein samples were resolved on 8% or 12% SDS-PAGE gels and were then transferred onto polyvinylidene difluoride (PVDF) membrane using either a wet or semi-dry transfer system. Membranes were blocked in 5% (w/v) dried milk in TBS-Tween (TBST) for at least 1 h at room temperature. Membranes were incubated with primary antibody, followed by the appropriate horseradish peroxidase-conjugated secondary antibody. They were developed using LumiGLO enhanced chemiluminescent (ECL) substrate (Cell Signalling). Bands were visualized using the GelDoc system (Biorad).

Quantitative PCR. Total RNA was isolated using the RNeasy Plus Mini kit (Qiagen) and quantified using a Nanodrop 2000 UV-visible spectrophotometer. cDNA was prepared using 20–100 ng μl⁻¹ total RNA by a reverse transcription PCR (RT-PCR) using a high capacity cDNA reverse transcription kit (Applied Biosystems), according to the manufacturer's instructions. Quantitative PCR (qPCR) was performed on cDNA using SYBR Green probes. qPCR was performed on a 7900 HT Fast Real-Time PCR System (Applied Biosystems) using Kapa fast master mix high ROX (Kapa Biosystems, for SYBR probes) or 2× PCR fast master mix (Applied Biosystems, for Taqman probes). For SYBR primer pair sequences, see Supplementary Methods. Fold changes in expression were calculated by the $\Delta\Delta C_t$ method using mouse *Rps18* as an endogenous control for mRNA expression. All fold changes are expressed normalized to the untreated control.

NQO1 bioassay. Inducer potency was quantified by use of the NQO1 bioassay in Hepa1c1c7 mouse hepatoma cells^{11,12}. Cells (10^4 per well of a 96-well plate) were grown for 24 h and exposed ($n = 8$) to serial dilutions of compounds for 48 h before lysis. NQO1 enzyme activity was quantified in cell lysates using menadione as a substrate. Protein concentrations were determined in aliquots from the same cell

lysates by the bicinchoninic acid (BCA) assay (Thermo Scientific). The CD value was used as a measure of inducer potency. For assays examining the effect of GSH on inducer potency, 50 μM of each compound was incubated with 1 mM GSH in the cell culture medium at 37°C for 30 min before treatment.

Preparation of rat liver mitochondria. Female Wistar rats aged between 10 and 12 weeks (Charles River) were culled by stunning and cervical dislocation before the liver being excised and stored in ice-cold buffer (STE buffer; 250 mM sucrose, 5 mM Tris-Cl, 1 mM EGTA (pH 7.4 at 4°C)). Rat liver mitochondria were isolated by homogenization and differential centrifugation at 4°C in STE buffer²⁶. In brief, minced tissue was homogenized in STE buffer before centrifugation (1,000g, 3 min, 4°C) and centrifuging the resulting supernatant (10,000g, 10 min, 4°C). The mitochondrial pellet was resuspended in fresh STE before centrifuging (10,000g, 10 min, 4°C). The resulting pellet was resuspended in STE and assayed for protein concentration via BCA assay (Thermo Scientific) against a BSA standard curve.

Preparation of bovine heart mitochondrial membranes. Bovine heart mitochondria were isolated by differential centrifugation in 250 mM sucrose, 10 mM Tris-Cl, 0.2 mM EDTA (pH 7.8 at 4°C). To prepare membranes, bovine heart mitochondria were blended with MilliQ water at 4°C before adding KCl to a final concentration of 150 mM and blending until homogenous. The suspension was centrifuged (13,500g, 40 min, 4°C) and the pellet was resuspended in re-suspension buffer (20 mM Tris-Cl, 1 mM EDTA, 10% glycerol, pH 7.55 at 4°C) before homogenization and assaying for protein by BCA assay (Thermo Scientific)²⁷.

Measuring complex II and III activity. Bovine heart mitochondrial membranes (80 μg protein per ml) were incubated in 50 mM potassium phosphate buffer (50 mM potassium phosphate, 1 mM EDTA, pH 7.4, 4°C) supplemented with 3 mM KCN, 4 μM rotenone and succinate. In a 96-well microplate, inhibitor or vehicle control and membrane incubation were plated and incubated for 10 min at 30°C. Alternatively, where indicated, itaconate was incubated with membranes and removed by twice centrifuging membranes and resuspending in non-itaconate containing buffer, before plating with 1 mM succinate. Oxidized cytochrome-c was added before measuring the respiratory chain activity by assessing the reduction of cytochrome-c spectrophotometrically at 550 nm at 20 s intervals for 5 min at 30°C. Final concentrations were 10 μg protein per well bovine heart membranes and 30 μM ferricytochrome c.

Measuring rat liver mitochondrial respiration. Respiration of rat liver mitochondria was assessed with an Oxygraph-2K (OROBOROS instruments high resolution respirometry). Rat liver mitochondria (0.5 mg mitochondrial protein per ml) were added to KCl buffer (pH 7.2, 37°C) and respiration assessed in the presence of 4 μg ml⁻¹ rotenone, 1 mM succinate, 1 μM FCCP and inhibitors or buffer control. **Assessing itaconate ester reactivity with glutathione.** GSH (1 or 5 mM) and 5 mM itaconate esters or vehicle control were incubated in KCl buffer (pH 7.2 or 8) at 37°C for 2 h, where indicated, 10 μg recombinant GST was added to the incubation. The reaction was stopped by acidification with 5% sulfosalicylic acid before assessing glutathione content by the GSH recycling assay as described previously²⁸.

Itaconate transport assays. Itaconate transport by mitochondrial carriers was assessed as described previously²⁹. For further details see Supplementary Methods.

Cell uptake of itaconate. C2C12 mouse myoblasts were plated at 300,000 cells per well in a 6-well plate in complete growth medium and adhered overnight in a humidified 5% CO₂, 37°C incubator. The following day, media was replaced with serum-free DMEM containing itaconate esters and cells were treated for 30 min at 37°C. Cells were extracted as described above (method for succinate quantification), with MS internal standard (100 pmol) added and stored at -80°C before LC-MS/MS analysis. For further details, see Supplementary Methods.

LC-MS/MS analysis was performed using an LCMS-8060 mass spectrometer (Shimadzu) with a Nexera X2 UHPLC system (Shimadzu). For further details, see Supplementary Methods.

KEAP1 cysteine target validation. COS1 cells (2.5×10^5 per well) in 6-well plates were co-transfected (Lipofectamine 2000) with 0.8 μg of Nrf2-V5 and 1.6 μg of wild-type or Cys151S mutant KEAP1¹⁴, or 1.6 μg of pcDNA. Cells were grown for 21 h then treated with 20 or 100 μM OI, 5 μM sulforaphane or 0.1% acetonitrile (vehicle) for 3 h. Cells were washed in PBS and lysed in 200 μl of SDS-lysis buffer (50 mM Tris-HCl, pH 6.8, 2% (w/v) sodium dodecyl sulfate (SDS) and 10% (v/v) glycerol). Lysates were sonicated (20 s at 30% amplitude using Vibra-Cell ultrasonic processor, Sonic) and boiled (3 min), and dithiothreitol (DTT) and Bromophenol blue were added up to 0.1 M and 0.02% (w/v) final concentrations, respectively. Proteins (10 μg) were resolved on a gradient (4–12%) NuPAGE SDS gel, transferred onto nitrocellulose membranes, and immunoblotted with anti-KEAP1 (rat monoclonal, Merk Millipore, clone 144), anti-Nrf2 (rabbit monoclonal, CST), and anti-β-actin (mouse monoclonal, Sigma) antibodies. Horseradish peroxidase (HRP)- or IRDye-labelled secondary antibodies were used interchangeably, followed by either ECL detection or scanning using Odyssey imager (Li-COR).

ELISA. Cytokine concentrations in cell supernatants were measured using ELISA DuoSet kits for mouse IL-10 and TNF and human IFN- β and IL-1 β , according to the manufacturer's instructions. Cytokine concentrations in serum samples isolated from whole blood were measured using Quantikine ELISA kits for mouse or human IL-1 β , IFN- β , IL-10 and TNF. DuoSet and Quantikine kits were from R&D Systems. Optical density values were measured at a wavelength of 450 nm, using a FLUOstar Optima plate reader (BMG Labtech). Concentrations were calculated using a four-parameter fit curve.

FACS analysis of ROS. BMDMs were seeded at 0.5×10^6 cells per ml and treated as normal. Then 2 h before staining, 100% ethanol was added to the dead cell control well. Thirty minutes before the end of the stimulation, CellROX (5 μ M) was added directly into the cell culture medium. Supernatants of cells that were to be stained with Aqua Live/Dead were removed, and an Aqua Live/Dead dilution (1 ml; 1:1,000 in PBS) was added to each well. Cells were incubated in tinfoil at 37 °C for 30 min. Cells were washed with PSB, scraped in PBS (0.5 ml), and transferred to polypropylene FACS tubes. Samples were analysed using a DAKO CyAn flow cytometer, and data was analysed using FlowJo software. MFI was quantified as a measure of cellular ROS production.

Nitric oxide assay. Nitric oxide concentrations in cell supernatants were measured using Greiss reagent assay kit from Thermo Fischer Scientific according to the manufacturer's instructions. Optical density values were measured at a wavelength of 548 nm, using a SoftMax Pro plate reader. Concentrations were calculated using a linear standard curve.

GSH/GSSG measurements. BMDMs were plated at 0.1×10^6 cells per ml in opaque 96-well plates. Cells were pre-treated with OI (125 μ M) for 2 h and then stimulated with hydrogen peroxide (100 μ M) for 24 h. After 24 h, cell media was removed and the reduced glutathione to oxidized glutathione (GSH/GSSG) ratio was quantified using MyBio GSH/GSSG-Glo Assay (V6611) as per manufacturer's instructions. Luminescence was quantified using a FLUOstar Optima plate reader.

LDH assay. Cells were plated at 0.5×10^6 cells per ml in white 24-well plates (500 μ l per well) and treated as required. Cytotoxicity, as determined by LDH release, was assayed using CytoTox96 Non-radioactive Cytotoxicity Assay kit (Promega) according to the manufacturer's instructions.

Seahorse analysis of lactate production. Cells were plated at 0.2×10^6 cells per well of a 24-well Seahorse plate. Cells were treated and stimulated as normal. A utility plate containing calibrant solution (1 ml per well) was placed in a CO₂-free incubator at 37 °C overnight. The next day, media was removed from cells and replaced with glucose-supplemented XF assay buffer (500 μ l per well) was placed in a CO₂-free incubator for at least 0.5 h. Compounds (glucose, oligomycin and 2-deoxy-D-glucose (2DG); 70 μ l) were added to the appropriate port of the injector plate. This plate together with the utility plate was run on the Seahorse for calibration. Once complete, the utility plate was replaced with the cell culture plate and run on the Seahorse XF-24.

Endotoxin-induced model of sepsis. For cytokine measurements, mice were treated intraperitoneally with OI (50 mg kg⁻¹) in 40% cyclodextrin in PBS or vehicle control for 2 h before stimulation with LPS (Sigma; 2.5 mg kg⁻¹) intraperitoneally for 2 h. Mice were euthanized in a CO₂ chamber, blood samples were collected and serum was isolated. Cytokines were measured using R&D ELISA kits according to manufacturer's protocol. For temperature recording, mice ($n = 10$ per group) were treated intraperitoneally with OI (50 mg kg⁻¹) in 40% cyclodextrin in PBS or vehicle control for 2 h before stimulation with LPS (5 mg kg⁻¹) and monitored for temperature at 1, 2, 3, 4, 6, 12, 18 and 24 h after LPS treatment. Temperature was monitored using subcutaneously implanted temperature transponder chips (Bio Medic Data Systems; IPTT 300) which were injected between the shoulder blades 48 h before experiment. At defined times, body temperature was measured by scanning the transponder with a corresponding BMDS Smart Probe. Animals were additionally monitored for clinical signs of endotoxin shock, based on temperature change, body condition, physical condition and unprovoked behaviour, with a combined score of 9 indicating the humane end point for the experiment.

siRNA transfection of BMDMs. Cells were plated at 1×10^6 cells per ml in 12-well plates overnight. On the day of transfection, the media was replaced with 500 μ l DMEM without penicillin/streptomycin or FBS. For each target gene, two Eppendorfs were prepared. OptiMax (250 μ l per well) was added to each tube. RNAiMax (add 5 μ l per well) was added to one set of tubes and short interfering siRNA (siRNA; 50 nM per well) was added to the second set of tubes. The tube containing the siRNA was added to the tube with RNAiMax, mixed well by pipetting and incubated for 15 min. The mix (500 μ l) was added to each well. Twenty-four hours after transfection, cells were treated as required.

Analysis of KEAP1 modification by OI. Human embryonic kidney cells (HEK293T cells) were transfected with a pCMV6-KEAP1 vector (Myc-DDK-tagged mouse KEAP1) (OriGene). C2C12, Hepa1c1c7 cells and COS1 cells

were from American Type Culture Collection (ATCC). The L929 cells are from Sigma (85011425). HEK293T cells were obtained from the Centre for Applied Microbiology and Research. Cell lines have not been tested for mycoplasma contamination. Twenty-four hours after transfection, cells were treated with OI (500 μ M) or vehicle control (PBS) for 4 h. Tagged KEAP1 was immunoprecipitated using an anti-Flag antibody (Sigma) and protein A/G beads (Santa Cruz). After immunoprecipitation, bound KEAP1 was eluted off the beads using Flag peptide (500 μ l; 200 μ g ml⁻¹) (Sigma) diluted in 1 \times TBS pH 7.4. The samples were then concentrated and the Flag peptide was removed using 10K centrifugation filter columns (Merck). The concentrated samples were then divided in half for downstream processing. One-half of each sample was diluted 1:2 with 5 \times SDS sample buffer and separated using SDS-PAGE (Bio-Rad). Overexpressed KEAP1 was detected using Coomassie blue staining and the corresponding bands were excised from the gel and subjected to in-gel digest as described. In brief, the gel slices were cut into smaller pieces (1–2 mm²) before reduction with DTT (10 mM) and alkylation with iodoacetamide (50 mM). Half of the gel slices from each sample were then subjected to a trypsin (2 μ g) digest, the other half were digested with elastase (1 μ g) overnight at 37 °C. Similarly, the remaining sample concentrates (in solution) were reduced with DTT and alkylated with iodoacetamide, before precipitation of the protein via the methanol-chloroform extraction method. The protein pellet was re-suspended in urea (6 M), which was then diluted to <1 M urea with ultrapure H₂O. The samples were then digested with trypsin (2 μ g) overnight at 37 °C. Digested protein samples were analysed in an Orbitrap Fusion Lumos coupled to a UPLC ultimate 3000 RSLCnano System (both Thermo Fisher). For further details, see Supplementary Methods.

Assessment of cysteine alkylation by itaconate using Iodo-TMT. After treatment, cells were lysed in HEPES pH 7.5, EDTA, glycerol and NP40. 2 mM TCEP and 50 mM NEM were added in a buffer containing 50 mM HEPES, 2% SDS, 125 mM NaCl, pH 7.2, and samples were incubated for 60 min at 37 °C in the dark to reduce and alkylate all unmodified protein cysteine residues. 20% (v/v) TCA was added to stabilize thiols and incubated overnight at 4 °C and then pelleted for 10 min at 4,000g at 4 °C. The pellet was washed three times with cold methanol (2 ml) and then resuspended in 2 ml 8 M urea containing 50 mM HEPES, pH 8.5. Protein concentrations were measured by BCA assay (Thermo Scientific) before protease digestion. Protein lysates were diluted to 4 M urea and digested with LysC (Wako) in a 1:100 enzyme:protein ratio and trypsin (Promega) at a final 1:200 enzyme:protein ratio for 4 h at 37 °C. Protein extracts were diluted further to a 2.0 M urea and LysC (Wako) at 1:100 enzyme:protein ratio and trypsin (Promega) at a final 1:200 enzyme:protein ratio were added again and incubated overnight at 37 °C. Protein extracts were diluted further to a 1.0 M urea concentration, and trypsin (Promega) was added to a final 1:200 enzyme:protein ratio for 6 h at 37 °C. Digests were acidified with 250 μ l of 25% acetic acid to a pH value of ~2, and subjected to C18 solid-phase extraction (50 mg Sep-Pak, Waters). Excess TMT label (6–7 M) was added to each digest for 30 min at room temperature (repeated twice). The reaction was quenched using 4 μ l 5% hydroxylamine. Samples were subjected to an additional C18 solid-phase extraction (50 mg Sep-Pak). For LC-MS/MS parameters, data processing and MS2 spectra assignment, TMT reporter ion intensities and quantitative data analysis, see Supplementary Methods.

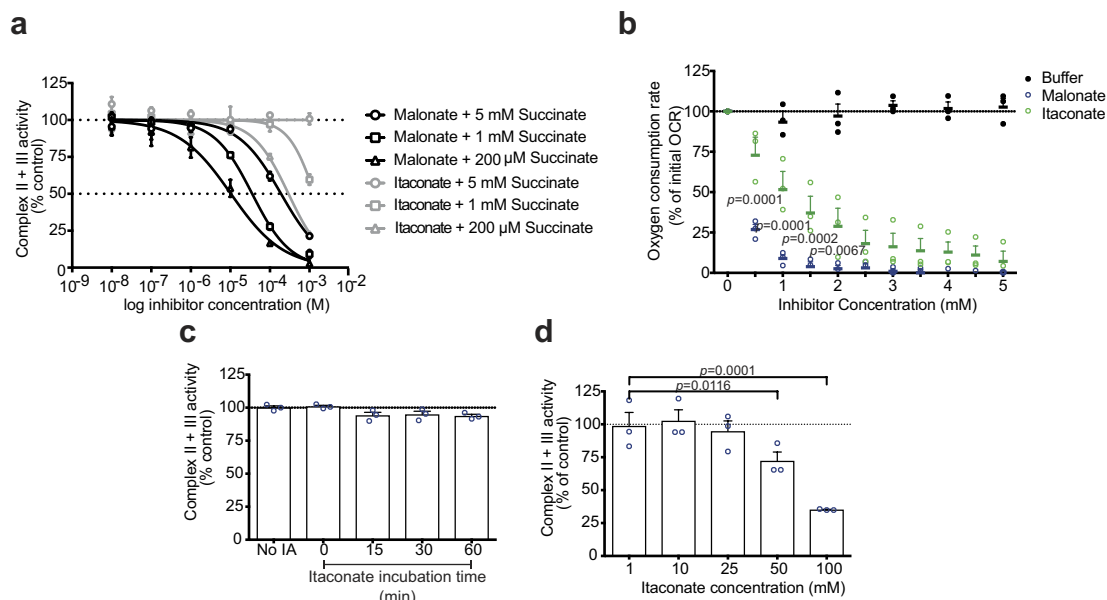
Reagents. For a complete list of reagents, see Supplementary Methods.

Mouse strains. Wild-type C57BL/6 mice were from Harlan UK and Harlan Netherlands. Animals were maintained under specific pathogen-free conditions in line with Irish and European Union regulations. Experiments were approved by local ethical review and were carried out under the authority of Ireland's project license. All animal studies performed in GSK were ethically reviewed and carried out in accordance with Animals (Scientific Procedures) Act 1986 and the GSK Policy on the Care, Welfare and Treatment of Animals. Nrf2-deficient mice and their wild-type counterparts, both on the C57BL/6 genetic background (used for isolation of BMDM cells), were bred and maintained in the Medical School Resource Unit of the University of Dundee.

Statistical analysis. Data were expressed as mean \pm s.e.m. and *P* values were calculated using two-tailed Student's *t*-test for pairwise comparison of variables, one-way ANOVA for multiple comparison of variables, and two-way ANOVA involving two independent variables. A Sidak's multiple comparisons test was used. A confidence interval of 95% was used for all statistical tests. Sample sizes were determined on the basis of previous experiments using similar methodologies. For all experiments, all stated replicates are biological replicates. For *in vivo* studies, mice were randomly assigned to treatment groups. For mass spectrometry analyses, samples were processed in random order and experimenters were blinded to experimental conditions.

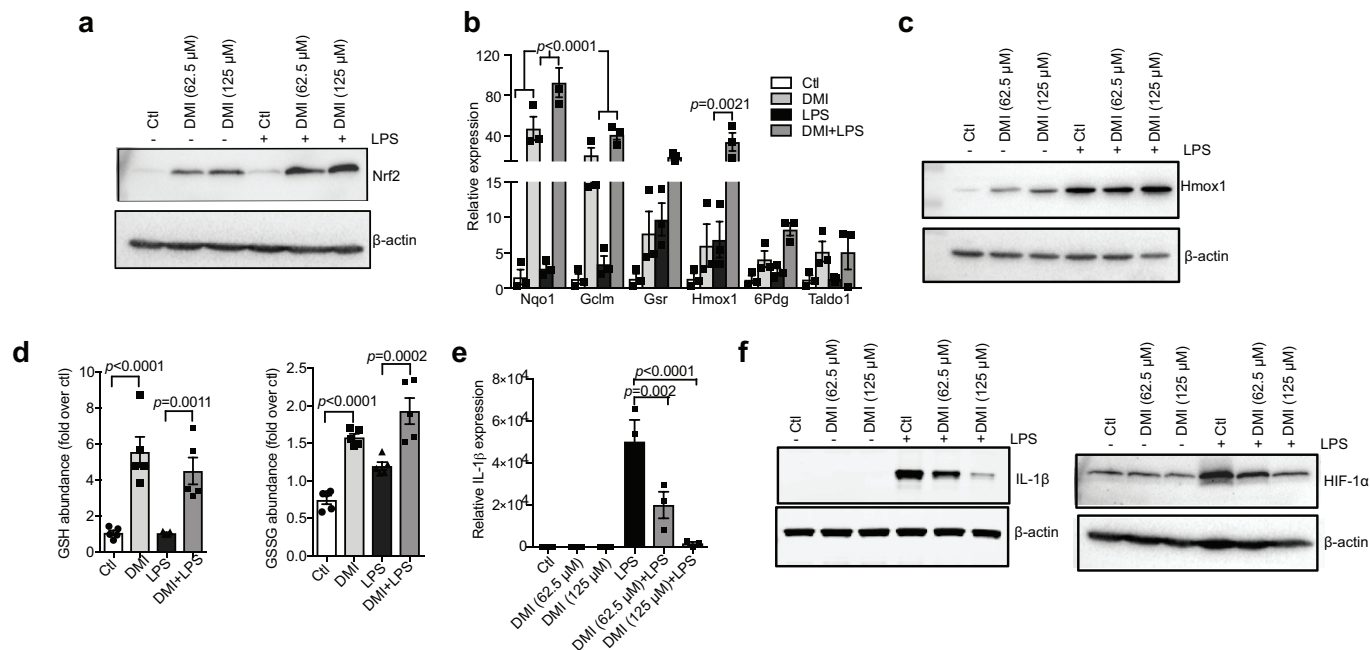
Data availability. Full scans for all western blots are provided in Supplementary Fig 1. Source Data for all mouse experiments have been provided. All other data are available from the corresponding author on reasonable request.

26. Chappell, J. B. & Hansford, R. V. A. *Subcellular Components: Preparation and Fractionation* 2nd edn (Butterworth, 1972).
27. Bridges, H. R., Mohammed, K., Harbour, M. E. & Hirst, J. Subunit NDUFV3 is present in two distinct isoforms in mammalian complex I. *Biochim. Biophys. Acta* **1858**, 197–207 (2017).
28. Akerboom, T. P. & Sies, H. Assay of glutathione, glutathione disulfide, and glutathione mixed disulfides in biological samples. *Methods Enzymol.* **77**, 373–382 (1981).
29. Booty, L. M. *et al.* The mitochondrial dicarboxylate and 2-oxoglutarate carriers do not transport glutathione. *FEBS Lett.* **589**, 621–628 (2015).



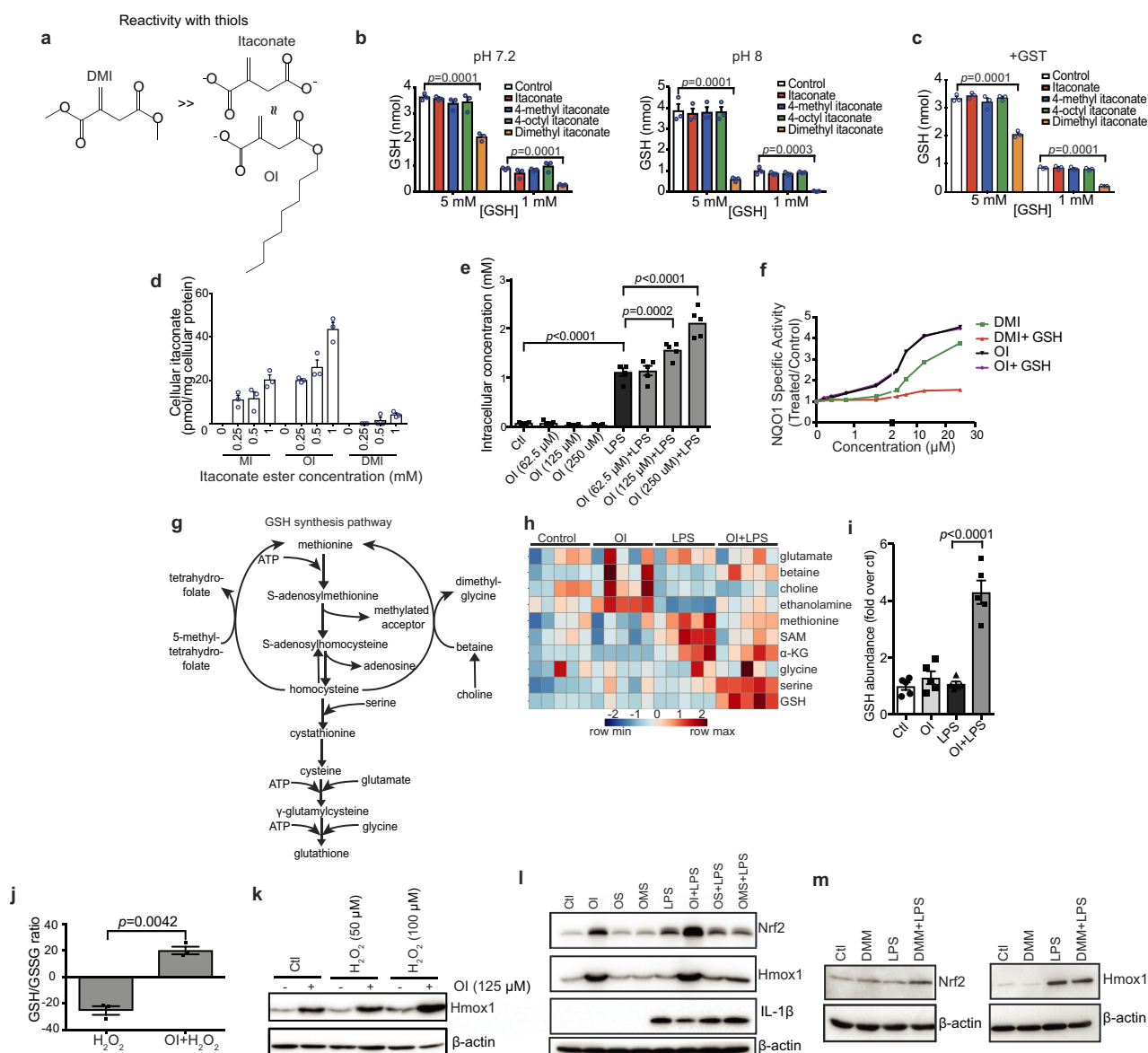
Extended Data Figure 1 | The effect of itaconate on complex II activity. **a**, Complex II and III activity in bovine heart mitochondrial membranes incubated with succinate plus malonate or itaconate ($n = 3$ independent experiments). **b**, Effect of malonate or itaconate on the oxygen consumption rate (OCR) of rat liver mitochondria in the presence of succinate (1 mM) and FCCP (1 μ M; $n = 3$ independent experiments).

c, d, Complex II and III activity in bovine heart mitochondrial membranes incubated with itaconate (IA; 1 mM unless indicated), with subsequent removal and addition of succinate (1 mM; $n = 3$ independent experiments) (see Methods for further details). Data are mean \pm s.e.m. P values calculated using one or two-way ANOVA.



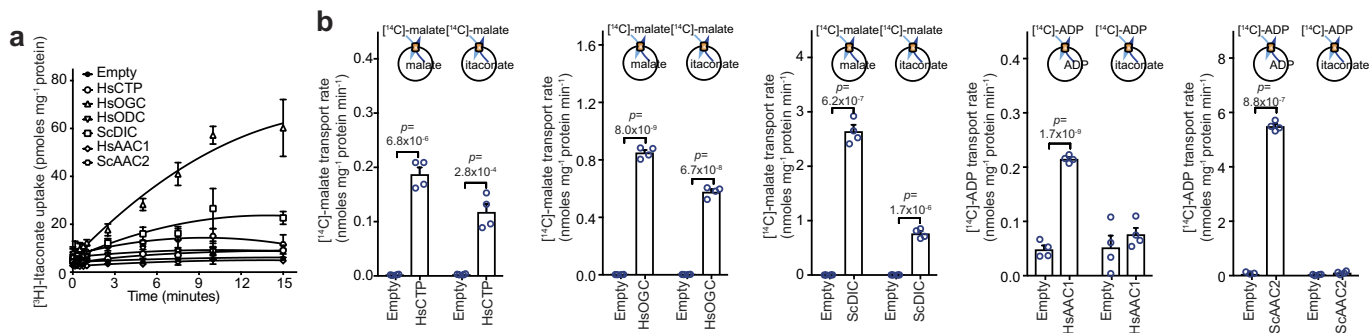
Extended Data Figure 2 | DMI activates Nrf2 and limits cytokine production. **a**, **c**, LPS (100 ng ml⁻¹)-induced Nrf2 (**a**, 24 h) and HMOX1 (**c**, 6 h) protein expression with or without the itaconate derivative DMI. **b**, Nrf2-dependent mRNA expression after treatment with LPS (6 h) and DMI where indicated ($n = 3$). **d**, Reduced glutathione (GSH) and oxidized glutathione (GSSG) levels after treatment with LPS and DMI

($n = 5$). **e**, **f**, LPS (24 h)-induced *Il1b* mRNA (**e**), IL-1 β and HIF-1 α protein (**f**) expression in mouse macrophages with or without DMI ($n = 3$). Data are mean \pm s.e.m. P values calculated using one-way ANOVA. Blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.



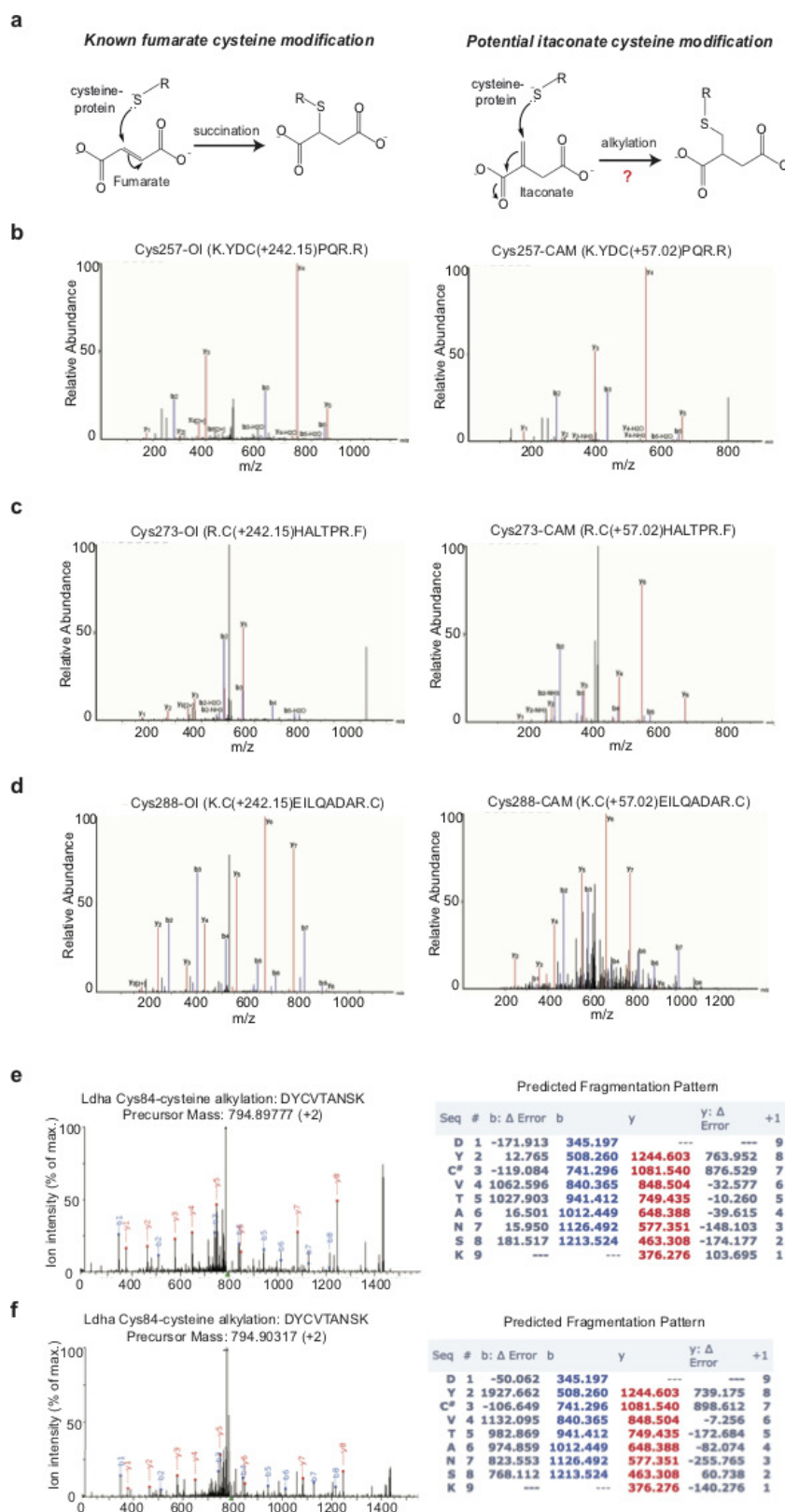
Extended Data Figure 3 | OI is the best tool to assess itaconate-dependent Nrf2 activity. **a**, Reactivity of DMI, itaconate and OI with thiols. **b**, **c**, Itaconate ester reactivity with GSH and glutathione-S-transferase (GST) as detailed in the Methods ($n = 3$). **d**, Itaconate levels in mouse C2C12 cells plus itaconate esters ($n = 3$). MI, 4-methyl itaconate. **e**, **i**, Itaconate (**e**) or GSH (**i**) levels plus LPS (6 h) and OI as indicated ($n = 5$). **f**, NQO1 activity in mouse Hepa1c7 cells treated with DMI or OI (48 h) and GSH ($n = 8$). **g**, **h**, Metabolic intermediates in GSH synthesis (**h**, average of five biological replicates). **i**, GSH levels after treatment with LPS (6 h) and/or OI ($n = 5$). **j**, GSH/GSSG ratio after treatment

with OI (2 h) and H_2O_2 (100 μM , 24 h; $n = 3$) as indicated. **k**, HMOX1 protein levels after treatment with OI and/or H_2O_2 (24 h). **l**, Nrf2, HMOX1 and IL-1 β protein levels in BMDMs pre-treated with OI, 4-octyl 2-methylsuccinate (OMS) or octyl succinate (OS), all 125 μM for 3 h with or without LPS (6 h). **m**, LPS-induced Nrf2 (24 h) and HMOX1 (6 h) protein expression with or without dimethyl malonate (DMM). Data are mean \pm s.e.m. P values calculated using one- or two-way ANOVA. Blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.



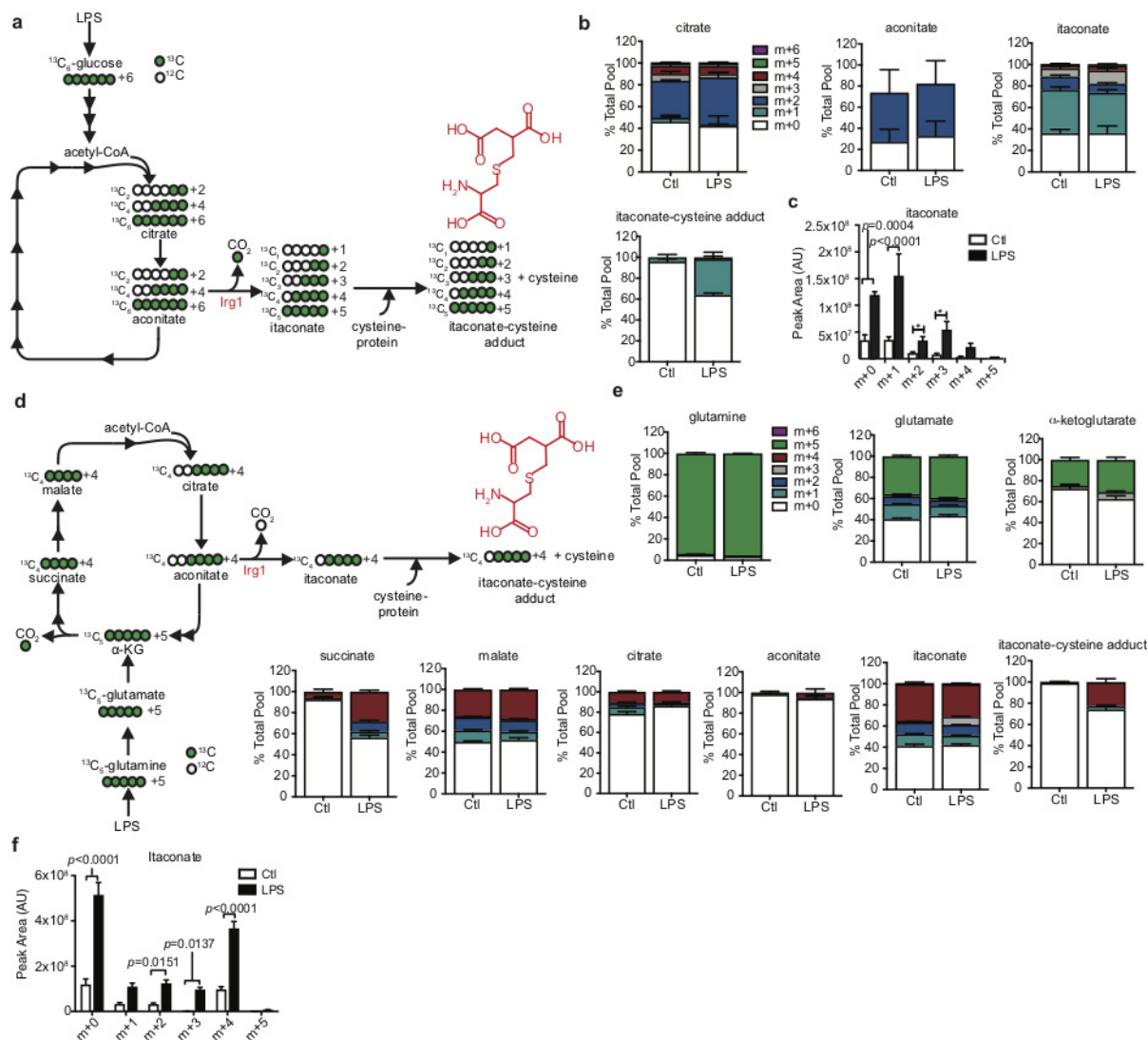
Extended Data Figure 4 | Itaconate is transported by the mitochondrial oxoglutarate, dicarboxylate and citrate carriers. a, Itaconate uptake into vesicles of *Lactococcus lactis* membranes expressing the indicated carriers loaded with itaconate (1 mM), and transport initiated by the addition of

[³H]itaconate (1 μM). **b**, Initial transport rates of each carrier with either canonical substrate (homo-exchange) or canonical substrate/itaconate (hetero-exchange). *n* = 4 independent experiments; data are mean ± s.d. *P* values calculated using two-tailed Student's *t*-test.



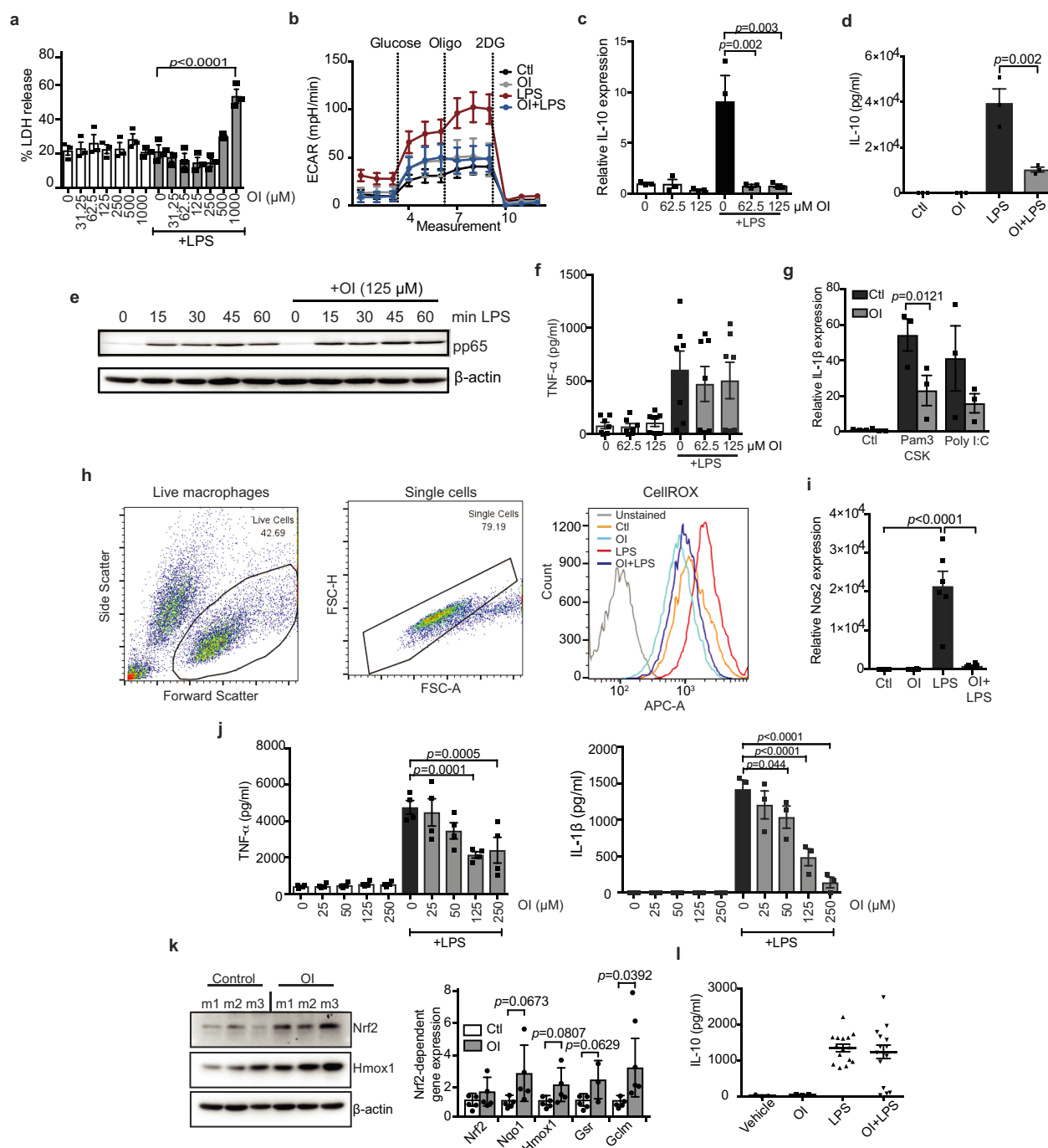
Extended Data Figure 5 | KEAP1 is alkylated by OI on major redox sensing cysteine residues. **a**, Modification of cysteine by fumarate or itaconate. Tandem mass spectrometry spectrum of KEAP1 Cys257 (**b**), Cys257 (**c**) and Cys288 (**d**) peptides, indicating alkylation of these sites after OI treatment (left) but not in the corresponding carbamidomethylated (CAM) peptides (right). **e**, **f**, LDHA Cys84

alkylation after treatment with LPS (**e**, 24 h) or OI (**f**, 250 μ M, 4 h) ($n = 4$). Detected N- and C-terminal fragment ions of both peptides are assigned in the spectrum and depicted as follows: *b*: N-terminal fragment ion; *y*: C-terminal fragment ion; asterisk: fragment ion minus NH_3 ; 0 or asterisk: fragment ion minus H_2O ; and 2+: doubly charged fragment ion. Representative of one independent experiment.



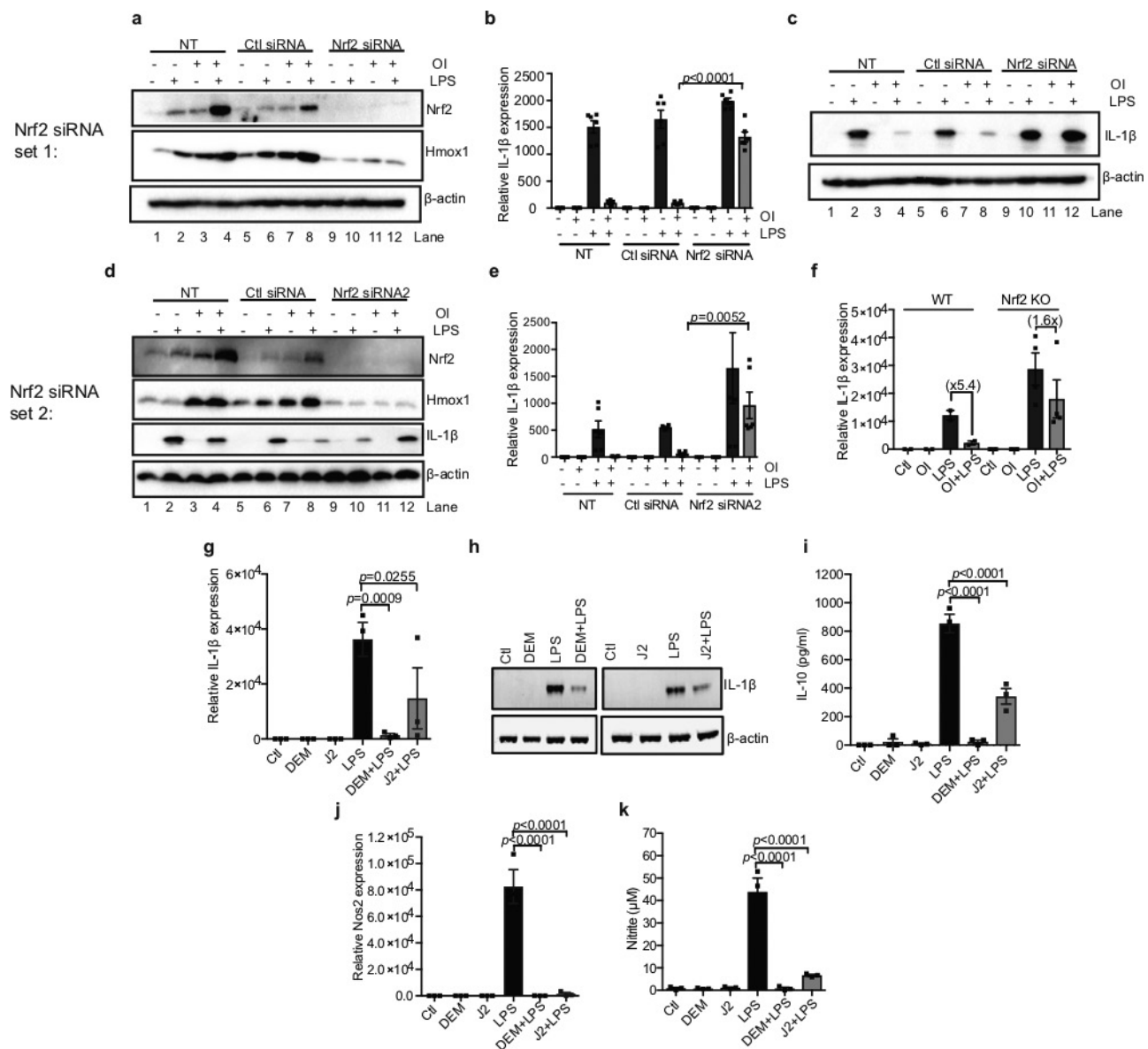
Extended Data Figure 6 | Identification of an itaconate-cysteine adduct. **a–c**, $^{13}\text{C}_6$ -glucose (**a–c**) or $^{13}\text{C}_5$ -glutamine (**d, e**) labelling experiment tracking itaconate-cysteine adduct formation in BMDMs treated with LPS ($n = 5$; 24 h). Data in **b** and **e** are expressed as the percentage

isotopologue of the total pool. Data in **c** and **f** represent changes in the total pool after LPS treatment. Data are mean \pm s.e.m., for five replicates. P values calculated using two-way ANOVA.



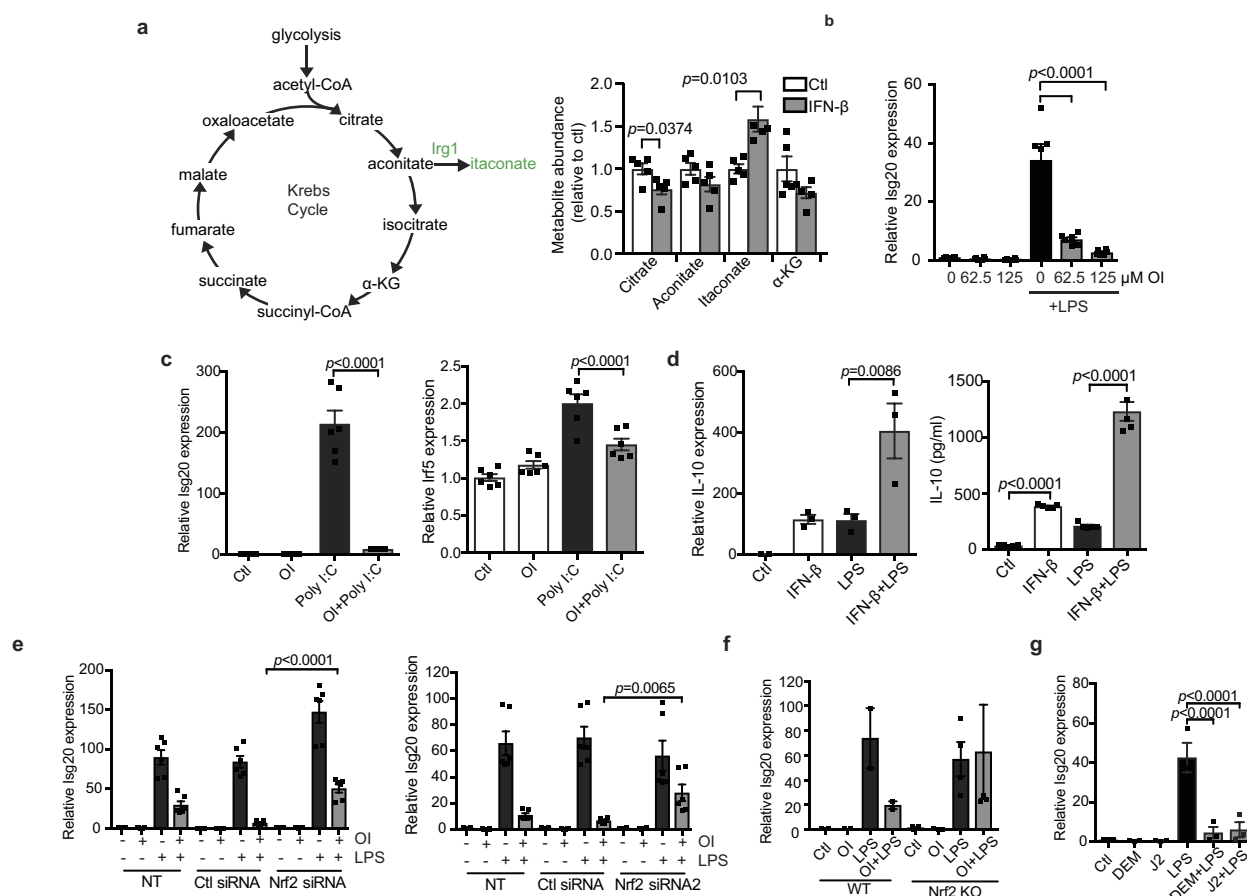
Extended Data Figure 7 | OI decreases LPS-induced cytokine production, extracellular acidification rate, ROS and nitric oxide. **a**, Percentage cytotoxicity (LDH release) in BMDMs after treatment with LPS and OI as indicated ($n = 3$). **b**, LPS-induced extracellular acidification rate (ECAR) after treatment with OI and/or LPS as indicated, analysed on the Seahorse XF-24 in BMDMs (trace representative of three independent experiments). **c**, **d**, LPS-induced *Il10* mRNA (**c**, 4 h) and protein (**d**, 24 h) and TNF protein (**f**; $n = 7$) after OI treatment as indicated ($n = 3$). **e**, Phosphorylated p65 (pp65) protein levels (a measure of NF- κ B activity) after treatment with LPS and OI as indicated. **h**, Representative gating strategy for FACS analysis of ROS production in cells as treated

in **d** (image representative of three independent experiments). **i**, LPS-induced NOS2 expression ($n = 6$), with or without OI treatment. **j**, LPS-induced TNF ($n = 4$) and IL-1 β ($n = 3$) protein levels after OI treatment in PBMCs. **k**, Nrf2 and HMOX1 protein levels or Nrf2-dependent gene expression ($n = 5$) in peritoneal macrophages from mice (m) injected intraperitoneally with OI (50 mg kg $^{-1}$, 6 h) or vehicle control. **l**, Serum IL-10 from mice injected intraperitoneally with vehicle control or OI (50 mg kg $^{-1}$, 2 h) and LPS (2.5 mg kg $^{-1}$, 2 h, $n = 3$ vehicle, OI; $n = 15$ LPS, OI plus LPS). Data are mean \pm s.e.m. P values calculated using one-way ANOVA. Blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.



Extended Data Figure 8 | The effects of OI on cytokine production are Nrf2-dependent. **a–e**, Nrf2, HMOX1 and IL-1 β protein levels (**a**, **c**, **d**) and *Il1b* mRNA expression (**b**, **e**) in mouse BMDMs transfected with two different *Nrf2* siRNAs (50 nM) compared with a non-silencing scrambled control siRNA plus LPS (6 h; **a–c**, **e**; 24 h; **d**) and/or OI ($n = 6$). NT, non-transfected. **f**, *Il1b* mRNA expression in wild-type and Nrf2-knockout BMDMs treated with LPS (24 h; WT $n = 2$, Nrf2 KO $n = 4$) and/or OI.

g–k, *Il1b* (**g**) and *Nos2* (**j**) mRNA, and IL-1 β (**h**), IL-10 (**i**), TNF and nitrite (**k**) production with or without LPS (24 h), diethyl maleate (DEM; 100 μ M) or 15-deoxy- Δ 12,14-prostaglandin J2 (J2; 5 μ M) pre-treatment for 3 h ($n = 3$). Data are mean \pm s.e.m. P values calculated using one-way ANOVA. Blots are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.



Extended Data Figure 9 | An Nrf2-dependent feedback loop exists between itaconate and IFN- β . **a**, Metabolite levels after treatment with IFN- β (1,000 U ml $^{-1}$; 27 h; $n=5$). **b**, **c**, *Isg20* and *Irf5* mRNA expression in BMDMs treated with LPS (**b**) or poly(I:C) (**c**, 40 μ g ml $^{-1}$; 24 h) and/or OI ($n=6$). **d**, *Il10* mRNA ($n=3$) and IL-10 protein ($n=5$) expression after treatment with LPS for 4 h (left) or 24 h (right) and/or IFN- β treatment (1,000 U ml $^{-1}$) for 3 h. **e**, *Isg20* expression in BMDMs

transfected with two different *Nrf2* siRNAs (50 nM) compared with non-silencing control plus LPS (6 h) and/or OI ($n=6$). **f**, *Isg20* mRNA expression in wild-type ($n=2$) and Nrf2-knockout ($n=4$) BMDMs plus LPS (6 h) and/or OI. **g**, *Isg20* mRNA expression after pre-treatment with LPS (24 h) and/or diethyl maleate (100 μ M) or 15-deoxy- Δ 12,14-prostaglandin J2 (5 μ M) for 3 h ($n=3$). Data are mean \pm s.e.m. *P* values calculated using one-way ANOVA.

Extended Data Table 1 | Mass spectrometry analysis of itaconate-induced cysteine alkylation

a

4-OI Residue	Peptide Amino Acid Position	Peptide Sequence	-10logP	Enzyme	Digest Type
Cys23	22-31	S.KC(+242.15)PEGAGDAV.M	31.47	Elastase	In Gel
Cys151	144-152	A.SISVGEKC(+242.15)V.L	44.11	Elastase	In Gel
	146-152	I.SVGEKC(+242.15)V.L	30.95		
	144-153	A.SISVGEKC(+242.15)V.L.H	30.25		
	145-152	S.ISVGEKC(+242.15)V.L	29.32		
Cys257	254-260	V.KYDC(+242.15)PQR.R	35.44	Elastase	In Gel
	255-260	K.YDC(+242.15)PQR.R	40.13	Trypsin	In Gel
	255-260	K.YDC(+242.15)PQR.R	41.08	Trypsin	In Solution
Cys273	273-279	R.C(+242.15)HALTPR.F	35.93	Trypsin	In Gel
	273-279	R.C(+242.15)HALTPR.F	38.65	Trypsin	In Solution
Cys288	282-293	L.QTQLQKC(+242.15)EILQA.D	41.70	Elastase	In Gel
	282-290	L.QTQLQKC(+242.15)EI.L	36.70		
	284-293	T.QLQKC(+242.15)EILQA.D	33.47		
	280-296	R.FLQTQLQKC(+242.15)EILQADAR.C	55.81		
	288-296	K.C(+242.15)EILQADAR.C	50.84		
	288-296	K.C(+242.15)EILQADAR.C	48.80		
Cys297	294-304	A.DARC(+242.15)KDYLVQI.F	37.15	Elastase	In Gel
K615	602-615	R.SGVGVATMEPCRK(+242.15).Q	37.59	Trypsin	In Gel
	602-615	R.SGVGVATM(+15.99)EPCRK(+242.15).Q	36.40		

b

Protein	Alkylated residue	Peptide amino acid position	Peptide sequence	X score	Ppm
Pls1	Cys111	97-123	KEGIC(+4.98)AIGGTSEQSSVGTQHSYSEEEK	5.998	-2.22
Acon	Cys385	378-395	VGLIGSC(+4.98)TNSSYEDMGR	4.212	4.17
Ldha	Cys84	82-90	DYC(+4.98)VTANSK	3.474	-2.91
Anxa1	Cys189	186-204	GDRC(+4.98)QDLSVNQDLADTDAR	3.514	-0.48
Ifi5b	Cys317	310-320	QMIEVPNC(+4.98)JTR	2.412	-4.16
Ipyr2	Cys156, 157	153-171	STDC(+4.98)C(+4.98)GDNDPIDVCEIGSK	4.664	22.60
Ef2	Cys41	33-42	STLTDSLVC(+4.98)K	3.677	-1.41
Thio	Cys73	73-81	C(+4.98)MPTFQFYK	2.422	-2.22

c

Protein	Alkylated residue	Peptide amino acid position	Peptide sequence	X score	Ppm
Gilt	Cys69	61-73	VSLYYESLC(+4.98)GACR	4.677	3.40
Fgd6	Cys1004	9996-1005	NVALLEQC(+4.98)K	3.759	-7.98
Olf644	Cys306	305-313	FC(+4.98)KILLGNK	3.155	-2.10
Ldha	Cys84	82-90	DYC(+4.98)VTANSK	2.832	3.88
Padi6	Cys553	553-558	C(+4.98)ISLNR	2.446	-18.58
Ubr4	Cys4241	4237-4244	LIASC(+4.98)HWK	2.421	-7.45
Hmx2	Cys314	314-323	C(+4.98)PFYAAQPK	2.279	2.89
Lhpp	Cys113	112-118	FC(+4.98)TNESQK	2.169	-11.64

a, Cysteine/lysine residue(s) in KEAP1 modified by OI as determined by tandem mass spectrometry. **b**, Cysteine residues modified by itaconate in BMDMs treated with LPS identified using tandem mass spectrometry. **c**, Cysteine residues modified by itaconate in BMDMs treated with OI identified using tandem mass spectrometry.

Structure of the peptidoglycan polymerase RodA resolved by evolutionary coupling analysis

Megan Sjodt¹, Kelly Brock², Genevieve Dobihal³, Patricia D. A. Rohs³, Anna G. Green², Thomas A. Hopf², Alexander J. Meeske³, Veerasak Srisuknimit⁴, Daniel Kahne⁴, Suzanne Walker³, Debora S. Marks², Thomas G. Bernhardt³, David Z. Rudner³ & Andrew C. Kruse¹

The shape, elongation, division and sporulation (SEDS) proteins are a large family of ubiquitous and essential transmembrane enzymes with critical roles in bacterial cell wall biology. The exact function of SEDS proteins was for a long time poorly understood, but recent work^{1–3} has revealed that the prototypical SEDS family member RodA is a peptidoglycan polymerase—a role previously attributed exclusively to members of the penicillin-binding protein family⁴. This discovery has made RodA and other SEDS proteins promising targets for the development of next-generation antibiotics. However, little is known regarding the molecular basis of SEDS activity, and no structural data are available for RodA or any homologue thereof. Here we report the crystal structure of *Thermus thermophilus* RodA at a resolution of 2.9 Å, determined using evolutionary covariance-based fold prediction to enable molecular replacement. The structure reveals a ten-pass transmembrane fold with large extracellular loops, one of which is partially disordered. The protein contains a highly conserved cavity in the transmembrane domain, reminiscent of ligand-binding sites in transmembrane receptors. Mutagenesis experiments in *Bacillus subtilis* and *Escherichia coli* show that perturbation of this cavity abolishes RodA function both *in vitro* and *in vivo*, indicating that this cavity is catalytically essential. These results provide a framework for understanding bacterial cell wall synthesis and SEDS protein function.

The synthesis of a cell wall and maintenance of its integrity are essential processes for virtually all Eubacteria, and the targeted disruption of cell wall biogenesis is among the most effective of therapeutic strategies in the treatment of bacterial infections. A central step in cell wall synthesis is the concatenation of a lipid II disaccharide pentapeptide headgroup onto a peptidoglycan chain through a glycosyl transfer reaction. This reaction has long been known to be catalysed by the glycosyltransferase domains of class A penicillin-binding proteins⁴ (aPBPs). However, deletion of all aPBPs is tolerated in *B. subtilis*⁵ and bacteria that lack aPBPs are able to synthesize peptidoglycan⁶, which implies the existence of non-aPBP glycosyltransferases. This was recently resolved with the discovery that the highly conserved SEDS membrane proteins comprise a second class of peptidoglycan polymerases^{1–3} (Fig. 1a). In fact, SEDS proteins are even more widely distributed than aPBPs^{1,7}. Despite their importance and broad phylogenetic distribution, however, SEDS protein function is not well understood and no SEDS protein has yet been characterized structurally.

To better understand SEDS protein function, we pursued structural studies of wild-type *T. thermophilus* RodA as well as a catalytically inactive RodA mutant (D255A). The proteins were expressed, purified and then crystallized using the lipidic cubic phase method. X-ray diffraction datasets were collected to a resolution of 2.9 Å and 3.2 Å for wild-type and D255A mutant proteins, respectively. Because RodA has no homologues of known structure, phase calculation by molecular replacement was impossible. A wide variety of heavy-atom phasing approaches were

attempted without success. Having exhausted conventional methods, we sought to develop an approach that requires neither experimental phase data nor the structure of a homologous protein.

Recent methodological advances in molecular replacement have expanded the range of suitable templates⁸, and evolutionary co-variation analysis now allows for fold prediction even in the absence of prior structural data^{9–11}. This approach exploits the fact that residues that interact with one another structurally tend to co-evolve to maintain their interactions. The analysis of many sequences enables inference of spatial interactions between pairs of residues, providing restraints sufficient to define major features of protein structure. Analysis of RodA by this method showed extensive co-variation throughout the protein (Fig. 1b, c), and we reasoned that using evolutionary coupling restraints to build RodA models might provide suitable templates for molecular replacement phasing, an approach we call ‘evolutionary coupling-enabled molecular replacement’ (EC-MR).

In brief, our EC-MR approach consisted of the parallel construction and sampling of many independent evolutionary coupling-derived models of RodA to identify suitable templates, followed by phase calculation and model building. First, 100 models of RodA were constructed on the basis of evolutionary restraints. These were each tested as single templates for molecular replacement in Phaser¹². Of these, 22 models yielded a cluster of solutions that were high-scoring and similar to one another (Fig. 1d). An ensemble search model was constructed from a subset of these, producing maps that were suitable for manual rebuilding followed by ROSETTA refinement in Phenix¹³. The final refined structure showed normal crystallographic statistics (Extended Data Fig. 1, Extended Data Table 1). While previous work has established that structural models constructed *ab initio* can in principle be suitable for molecular replacement phasing in select cases^{14,15}, these analyses were conducted on very short proteins (<100 amino acids) with high-resolution structural data (<2.1 Å). Other work has shown that at very high resolution, structures can be solved using even a single atom as a search model¹⁶. At lower resolution, symmetric α -helical proteins can be phased from helical fragments, although this relies on symmetry conditions that are met in only a small minority of cases¹⁷. Determination of the structure of RodA by EC-MR establishes that evolutionary covariance-derived models can be suitable for phase determination of even a large (359 amino acid) asymmetric protein at modest resolution.

The structures of wild-type RodA and the D255A mutant are virtually identical (Extended Data Fig. 2), and we focus here on the higher-resolution wild-type RodA structure. The overall structure shows ten well-resolved transmembrane helices (TM1–TM10) connected by loops, most of which are well ordered (Fig. 2). Searches for proteins of similar fold with the DALI server¹⁸ yielded no hits, indicating that RodA possesses a unique overall fold. The transmembrane helices of RodA are largely straight and perpendicular to the membrane

¹Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. ²Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. ³Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

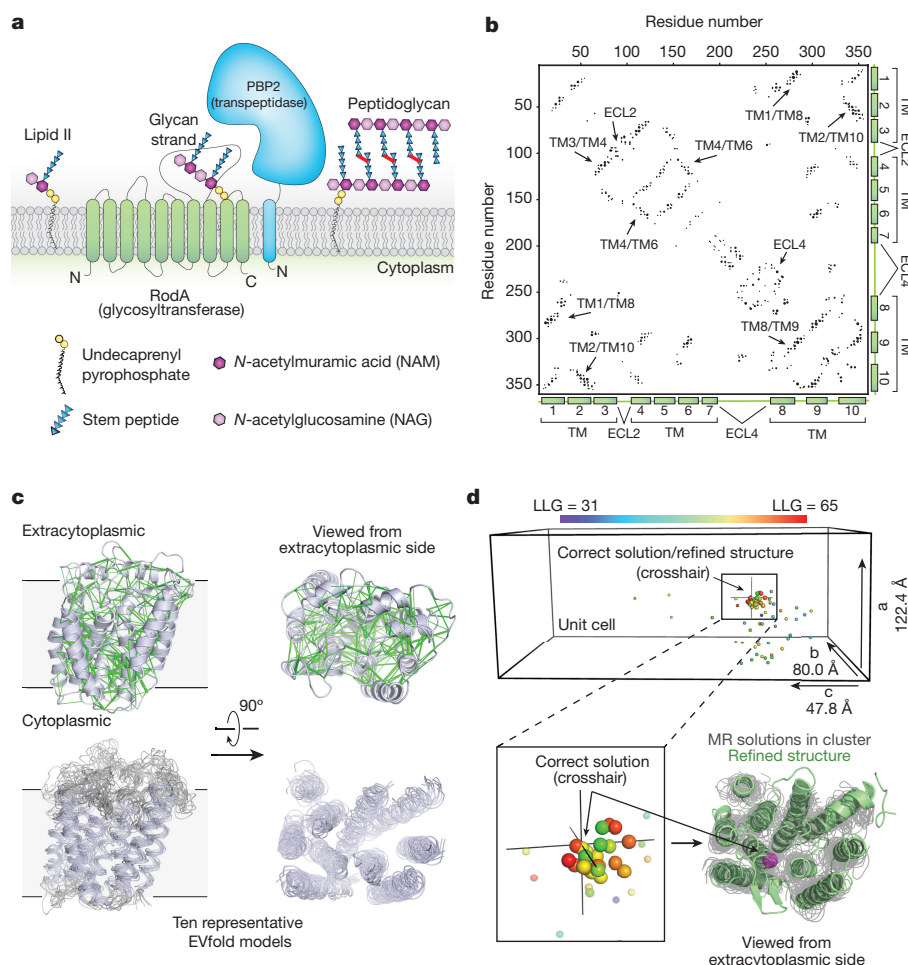


Figure 1 | Biological role of RodA and evolutionary co-variation fold prediction.

a, RodA is a peptidoglycan polymerase. **b**, Evolutionary co-variation map showing 476 co-evolved residues in RodA. **c**, Evolutionary couplings (green lines) were used to generate models of RodA. **d**, Molecular replacement (MR) solutions obtained from 100 RodA models. As a reference point, the C α atom of Glu108 is shown as a sphere coloured according to log-likelihood gain (LLG) score for each solution. The corresponding position for the final refined RodA structure is shown as a crosshair. The refined structure (green) and the C α atom of Glu108 (magenta sphere) is overlaid with the top MR solutions shown in grey.

plane, with the exception of TM3 that runs diagonally through the membrane with a 45° kink at Pro71. Although some previous studies have suggested that the SEDS protein FtsW could function as a lipid II flippase¹⁹, RodA lacks a transmembrane channel suitable for lipid II transport and bears no notable structural similarity to transporters or flippases.

The intracellular loops of RodA are structured but are very short, and little polar surface area is exposed on the intracellular face of the enzyme. By contrast, extracellular loops (ECLs) 2, 4 and 5 are large and contain many functionally essential residues¹, consistent with catalytic activity occurring at the extracytoplasmic face of the membrane. ECL2 includes a highly conserved β -hairpin, capped by Gly100 and Pro101. ECL4 is even larger at 80 amino acids in length, but is not resolved from residues 189 to 227 as well as from 237 to 251. These regions include essential residues as well as high-ranking evolutionary couplings between ECL4 and ECL1, ECL2 and ECL5, which suggests that they have functionally important roles despite not being resolved in the current structure (Extended Data Figs 3, 4). It is possible that these regions become ordered only upon substrate binding, or in complex with a peptidoglycan crosslinking enzyme. Unlike other ECLs, ECL5 is not exposed to the surface and instead is buried within the protein core.

Between TM2 and TM3 is a long hydrophobic groove containing electron density suggestive of a bound lipid molecule, which we tentatively modelled as monoolein owing to its high concentration in the crystallization conditions (Fig. 2b, c, Extended Data Figs 2, 5). This groove is adjacent to a collection of highly conserved residues (Fig. 2c), and may represent the binding site for the lipid-anchored substrates of RodA. Adjacent to this groove is a large water-filled cavity open to the extracellular face of the protein, flanked by Glu108, Met306, Leu307, Gln310 and Thr342 (Fig. 2d). On the edge of this cavity, Glu108 and

Lys111 form an absolutely conserved salt bridge (Fig. 2e). Mapping the results of previous high-throughput mutagenesis in *B. subtilis* RodA onto the structure shows that the salt bridge and other residues in the central cavity are intolerant of substitution¹ (Fig. 3a).

To investigate the function of this central cavity in more detail, we turned to site-directed mutagenesis followed by phenotypic characterization in the representative model Gram-negative and Gram-positive organisms *E. coli* and *B. subtilis*. First, we assessed whether the conserved salt bridge between Glu108 and Lys111 is essential, by monitoring the effect of the mutant enzyme when expressed in a wild-type background. Consistent with previous results, mutation of either of these residues to alanine resulted in a dominant-negative phenotype characteristic of Rod complex dysfunction, in which cells lose their elongated morphology and become enlarged and spherical before lysis (Fig. 3b). This suggests that the mutant proteins are properly folded and able to interact with other members of the Rod complex, but are unable to promote cell elongation.

To ascertain whether the salt bridge is catalytic or merely important for proper folding, we constructed a salt-bridge swap mutant (that is, an E108K and K111E double mutant). If the role of the salt bridge is purely structural, this swap should have a minimal effect on RodA function, because the salt bridge is maintained. However, if Glu108 has a role in catalysis, then swapping the two amino acids would be expected to abrogate function. Indeed, the mutant protein folds properly as measured by circular dichroism spectroscopy (Extended Data Fig. 6), but in cell assays the mutant protein shows a strong dominant-negative phenotype (Fig. 3b, c, Extended Data Fig. 7). Moreover, the mutant enzyme has no detectable peptidoglycan polymerization activity *in vitro*, confirming that its toxicity derives from a lack of catalytic activity (Fig. 3d, Supplementary Fig. 1). Other residues near the central cavity were similarly essential for RodA function, including Asp255

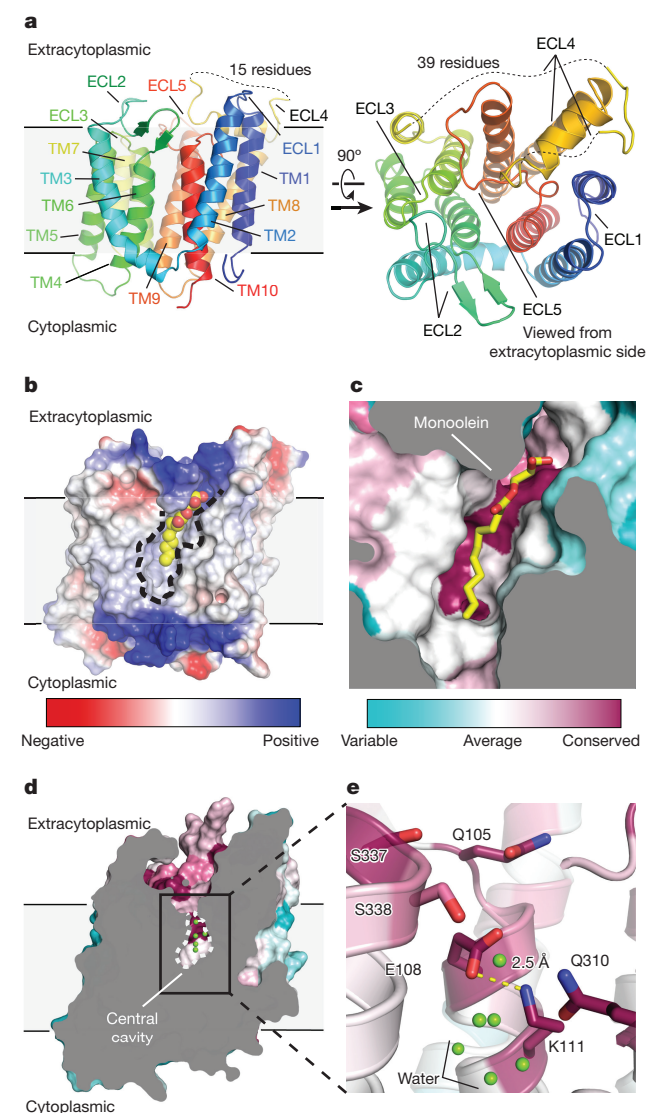


Figure 2 | Structure of RodA. **a**, Structure of RodA viewed parallel to the membrane plane (left) and from the extracytoplasmic side (right). **b**, Surface view of RodA showing electrostatic potential and the position of a bound lipid between TM2 and TM3 (orientation is identical to that in **a**). **c**, Close-up view of bound lipid (yellow). RodA surface is coloured by sequence conservation. **d**, In the centre of RodA there is a water-filled cavity open to the extracytoplasmic surface. **e**, The water-filled cavity of RodA is flanked by highly conserved polar residues and a salt bridge (yellow dashed line) between Glu108 and Lys111.

as previously reported¹, as well as Asp152 (Fig. 3b, Extended Data Fig. 7). Analysis of evolutionary co-variation data using EVmutation²⁰ likewise predicts these residues, and the salt bridge, to be immutable. Taken together, the high degree of sequence conservation, intolerance to mutation and catalytic essentiality of residues surrounding the central cavity confirm that this portion of the protein has a critical role in peptidoglycan polymerization, which makes it a prime target for the development of antibiotics.

The glycan-strand polymerization process catalysed by RodA and aPBPs is essential but not sufficient to build a cell wall. A second key step is peptide crosslinking, catalysed by the penicillin-binding domains of aPBPs and class B penicillin-binding proteins (bPBPs). Cytological and protein–protein interaction studies indicate that SEDS proteins and bPBPs are likely to form a complex in cells^{1,2}, and evolutionary coupling analysis shows strong co-variation between bBP and SEDS protein sequences. This is sufficient to map the binding site

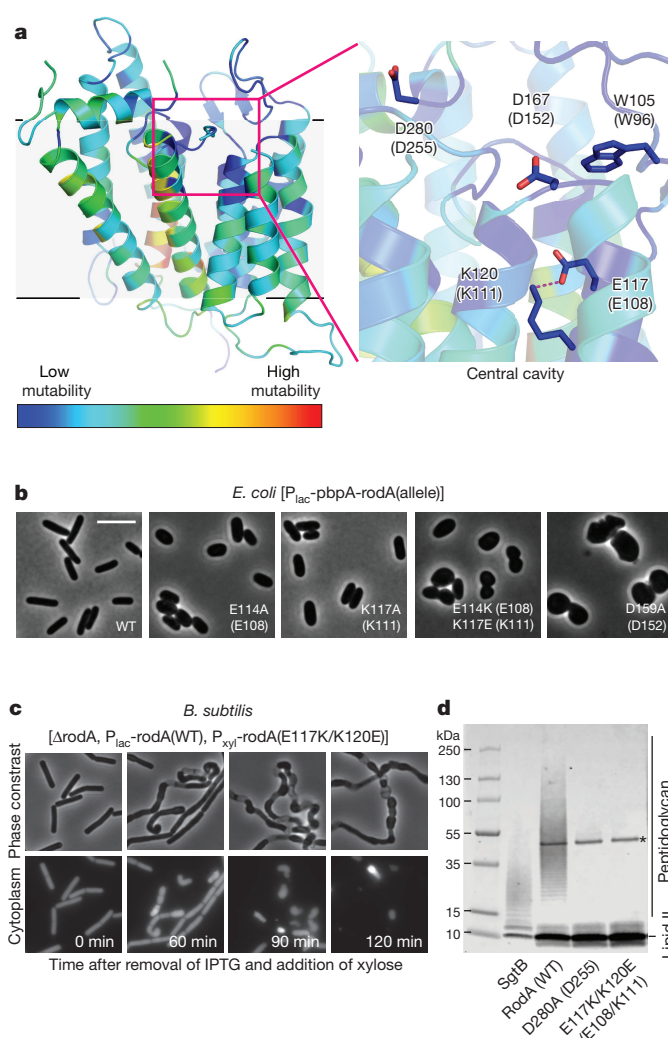


Figure 3 | The central cavity is essential for RodA function. **a**, Homology model of *B. subtilis* RodA with each residue coloured according to its tolerance for mutation (see Methods). Residues are numbered according to *B. subtilis*, with corresponding *T. thermophilus* numbering in parentheses. **b**, Wild-type *E. coli* cells harbouring a plasmid with indicated RodA mutants. Scale bar, 5 μm. WT, wild type. **c**, Expression of RodA charge-swap mutant is toxic in *B. subtilis*. Intracellular mCherry expression indicates cytosol. For both **b** and **c**, images are representative of three independent experiments. IPTG, isopropyl-β-D-thiogalactoside. **d**, Catalytic competence of *B. subtilis* RodA mutants, representative of two independent experiments. Asterisk indicates the PBP4 labelling enzyme used for detection; SgtB, SgtB(Y181D) used as control.

between bPBPs and RodA to TM8 and TM9 (Fig. 4a, b, Extended Data Fig. 8), corresponding to the proposed interaction site between the divisome SEDS protein FtsW and its corresponding bBP, FtsI^{10,21,22}. RodA mutants in this interface exhibit a dominant-negative effect in *E. coli* (Fig. 4c). However, mutation of this site does not prevent RodA-mediated peptidoglycan polymerization *in vitro* (Extended Data Fig. 7), confirming the functional importance of coordinated glycan strand elongation and peptide crosslinking in cells.

Complexes between SEDS proteins and bPBPs contain both glycan-strand polymerization and peptide crosslinking active sites, recapitulating the dual catalytic activities found in aPBPs (Fig. 4d). Understanding how SEDS proteins coordinate their activity with that of bPBPs to build a cell wall will be an important area for future investigation. The structure of RodA now provides a foundation for such work, serving as a framework for understanding the function of SEDS proteins.

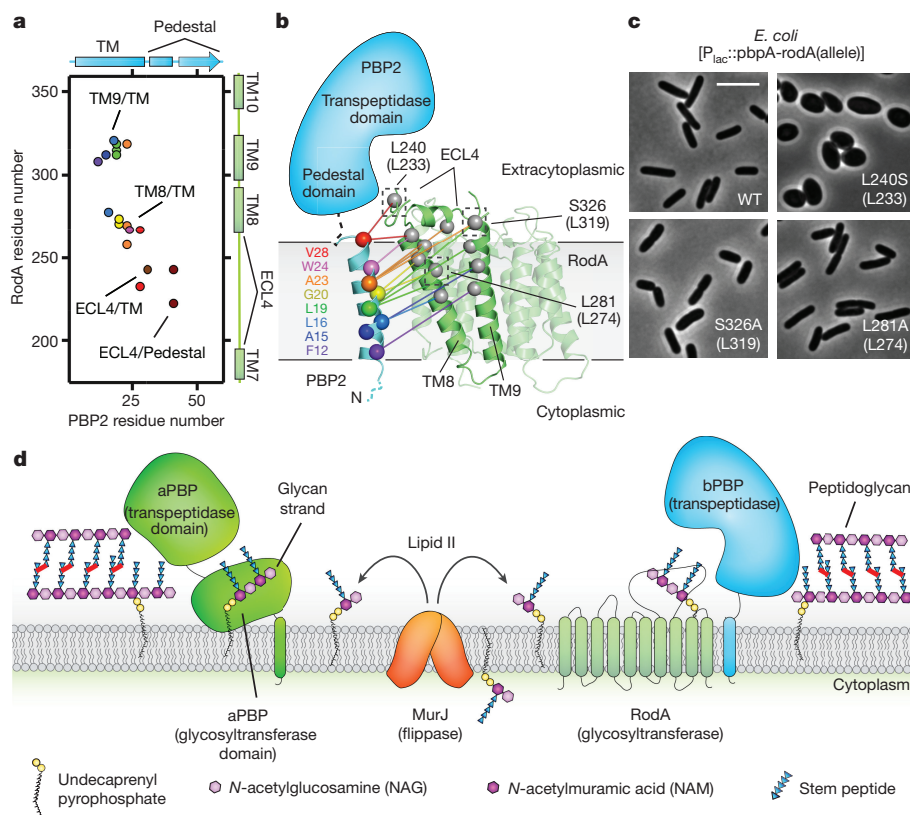


Figure 4 | Interaction between RodA and its class B penicillin-binding protein, PBP2. **a**, Evolutionary co-variation map showing 19 evolutionary couplings between RodA and PBP2. **b**, Representation of evolutionary co-variation. **c**, Mutations in RodA-bPBP interface result in morphological abnormalities. Mutations were made in *E. coli* RodA, with *T. thermophilus* numbering in parentheses. Data are representative of two independent experiments. Scale bar, 5 μ m. **d**, Model of peptidoglycan biogenesis. Lipid II is flipped across the membrane by MurJ, and polymerized and crosslinked by a complex of a SEDS protein (RodA) and bPBP, or by a bifunctional aPBP.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 October 2017; accepted 8 February 2018.

Published online 28 March 2018.

- Meeske, A. J. *et al.* SEDS proteins are a widespread family of bacterial cell wall polymerases. *Nature* **537**, 634–638 (2016).
- Cho, H. *et al.* Bacterial cell wall biogenesis is mediated by SEDS and PBP polymerase families functioning semi-autonomously. *Nat. Microbiol.* **1**, 16172 (2016).
- Emami, K. *et al.* RodA as the missing glycosyltransferase in *Bacillus subtilis* and antibiotic discovery for the peptidoglycan polymerase pathway. *Nat. Microbiol.* **2**, 16253 (2017).
- Goffin, C. & Ghuysen, J.-M. Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. *Microbiol. Mol. Biol. Rev.* **62**, 1079–1093 (1998).
- McPherson, D. C. & Popham, D. L. Peptidoglycan synthesis in the absence of class A penicillin-binding proteins in *Bacillus subtilis*. *J. Bacteriol.* **185**, 1423–1431 (2003).
- Packiam, M., Weinrick, B., Jacobs, W. R., Jr & Maurelli, A. T. Structural characterization of mucopeptides from *Chlamydia trachomatis* peptidoglycan by mass spectrometry resolves ‘chlamydial anomaly’. *Proc. Natl Acad. Sci. USA* **112**, 11660–11665 (2015).
- Otten, C., Brilli, M., Vollmer, W., Viollier, P. H. & Salje, J. Peptidoglycan in obligate intracellular bacteria. *Mol. Microbiol.* **107**, 142–163 (2018).
- DiMaio, F. *et al.* Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473**, 540–543 (2011).
- Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
- Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).
- Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
- McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
- DiMaio, F. *et al.* Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **10**, 1102–1104 (2013).
- Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259–264 (2007).
- Rigden, D. J., Keegan, R. M. & Winn, M. D. Molecular replacement using *ab initio* polyaniline models generated with ROSETTA. *Acta Crystallogr. D* **64**, 1288–1291 (2008).

- McCoy, A. J. *et al.* *Ab initio* solution of macromolecular crystal structures without direct methods. *Proc. Natl Acad. Sci. USA* **114**, 3637–3641 (2017).
- Strop, P., Brzustowicz, M. R. & Brunger, A. T. *Ab initio* molecular-replacement phasing for symmetric helical membrane proteins. *Acta Crystallogr. D* **63**, 188–196 (2007).
- Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549 (2010).
- Mohammadi, T. *et al.* Identification of FtsW as a transporter of lipid-linked cell wall precursors across the membrane. *EMBO J.* **30**, 1425–1432 (2011).
- Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
- Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
- Leclercq, S. *et al.* Interplay between penicillin-binding proteins and SEDS proteins promotes bacterial cell wall synthesis. *Sci. Rep.* **7**, 43306 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support for the work was provided by NIH grant U19AI109764 (A.C.K., D.Z.R., T.G.B., S.W. and D. K.), NIH grant R01GM106303 (D.S.M.) and a CIHR doctoral research award to P.D.A.R. We thank Advanced Photon Source GM/CA beamline staff for technical support during X-ray data collection, K. Arnett (Harvard Center for Macromolecular Interactions) for support of circular dichroism experiments and C. Sander for discussions.

Author Contributions M.S. and A.J.M. performed expression screening experiments, and M.S. performed large-scale purification and crystallization of RodA as well as enzyme assays and circular dichroism spectroscopy. Additional input regarding enzyme assays was provided by P.D.A.R., V.S., D.K. and S.W. The structure was solved and refined by M.S. and A.C.K. using evolutionary coupling-derived models developed by K.B., A.G.G., T.A.H. and D.S.M. Assessment of RodA mutant phenotypes was conducted by G.D. and P.D.A.R. with supervision from T.G.B. and D.Z.R. Overall project supervision was performed by A.C.K. with input from T.G.B. and D.Z.R. The manuscript was written by M.S. and A.C.K. with input from other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.C.K. (andrew.kruse@hms.harvard.edu).

Reviewer Information Nature thanks R. Read, K. Young and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein purification. RodA from *T. thermophilus* was cloned into pAM172 plasmid¹ using EcoRI and AvrII restriction enzymes (resulting in plasmids pMS211 and pMS224 for wild-type and D255A RodA, respectively). The expression plasmids contain an amino-terminal SUMO-fusion followed by a Flag epitope tag and a 3C protease cleavage site (SUMO–Flag–3C–RodA and SUMO–Flag–3C–RodA(D255A)), and were transformed into *E. coli* C43 derivative of BL21 (DE3) harbouring an arabinose-inducible Ulp1 protease plasmid (pAM174) under the selection for both plasmids. Five fresh transformants harbouring both plasmids were inoculated into 5 ml LB medium supplemented with 100 µg ml^{−1} ampicillin and 35 µg ml^{−1} chloramphenicol and allowed to grow overnight at 37°C in a rolling shaker. The 5-ml overnight culture was then diluted in 1 litre of TB broth supplemented with 0.1% glucose, 2 mM MgCl₂, 100 µg ml^{−1} ampicillin, and 35 µg ml^{−1} chloramphenicol. Cultures were grown at 37°C until an OD₆₀₀ of 0.6 and shifted to 20°C. At an OD₆₀₀ of 0.8, protein expression was induced by addition of IPTG (1 mM final) and arabinose (0.2% final) for RodA and Ulp1, respectively. After a 16-h induction, cells were collected and frozen at −80°C.

Cells were resuspended in lysis buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 20 mM MgCl₂, 100 µg ml^{−1} lysozyme, 1:100,000 (v/v) benzonase nuclease, and 2 mg ml^{−1} iodoacetamide), lysed by sonication and membranes were collected by ultracentrifugation at 100,000g for 1 h at 4°C. Flag–3C–RodA was then extracted using a glass dounce tissue grinder in a solubilization buffer containing 20 mM HEPES pH 7.5, 500 mM NaCl, 20% (v/v) glycerol, 2 mg ml^{−1} iodoacetamide, and 1% (w/v) *N*-dodecyl β-D-maltoside (DDM; Anatrace). Samples were stirred for 2 h at 4°C, then centrifuged as above for 1 h. The supernatant containing solubilized RodA was supplemented with 2 mM CaCl₂ and loaded by gravity flow onto 4 ml anti-Flag antibody affinity resin. The resin was washed extensively, first in 50 ml of buffer containing 20 mM HEPES pH 7.0, 500 mM NaCl, 2 mM CaCl₂, 20% glycerol, 0.1% DDM, and then in 50 ml of the same buffer supplemented with 10 mM adenosine 5′-triphosphate magnesium salt and 20 mM KCl to remove the bacterial chaperones, GroEL and DnaK. RodA was eluted in 20 mM HEPES pH 7.0, 500 mM NaCl, 20% glycerol, 0.1% DDM supplemented with 5 mM EDTA and 0.2 mg ml^{−1} Flag peptide. 3C protease was added (1:1,000 w/w) and incubated with RodA at 4°C overnight. RodA was further purified by size exclusion chromatography (SEC) on a Sephadex S200 column (GE Healthcare) in buffer containing 20 mM HEPES pH 7.5, 500 mM NaCl and 0.1% DDM. After preparative SEC, the protein was concentrated to 30–40 mg ml^{−1} and flash-frozen with liquid nitrogen in aliquots of 8 µl. Samples were stored at −80°C until use for crystallography. Purity and monodispersity of crystallographic samples was evaluated by SDS–PAGE and analytical SEC, respectively.

Crystallography and data collection. Purified *T. thermophilus* wild-type and D255A RodA were reconstituted into lipidic cubic phase by mixing with a 10:1 (w/w) mix of monoolein (Hampton Research) with cholesterol (Sigma Aldrich) at a ratio of 1.0:1.5 protein:lipid by mass, using the coupled syringe reconstitution method²³. All samples were mixed at least 100 times before dispensing. The resulting phase was dispensed in 15–40 nl drops onto a glass plate and overlaid with 600 nl of precipitant solution using a Gryphon LCP robot (Art Robbins Instruments). Crystals grew in precipitant solution containing 35–50% PEG 200, 100 mM NaCl, 100 mM MgCl₂ and 100 mM Tris pH 7.6–8.2. Initial crystallization hits grew within 24 h, with diffraction-quality crystals reaching full size over the course of 2–4 weeks. Crystals were collected using mesh loops and stored in liquid nitrogen until data collection. Data collection was carried out at Advanced Photon Source GM/CA beamline 23ID-B. An initial grid raster with 80 × 30-µm beam dimensions was performed to locate crystals within the loop. Additional fine-tuning rasters were performed using a 10-µm beam diameter to optimize the position of the crystal for data collection. Data were collected using a 10-µm beam and 0.2°-oscillation width per frame at a wavelength of 1.033 Å and a fivefold attenuation factor. For both wild-type and D255A RodA, a complete dataset was obtained from a single crystal. Diffraction data were indexed and processed using XDS²⁴. Both wild-type and D255A RodA crystallized in the C2 space group with one molecule in the asymmetric unit, with a solvent content of approximately 60%.

Generation of *ab initio* evolutionary coupling-derived models of RodA. Multiple sequence alignments (MSA) of the full length *T. thermophilus* RodA (Uniprot ID Q5SIX3) were generated using the iterative hidden Markov model-based sequence search tool jackhammer²⁵ with five iterations. Alignments were built using the Uniref100 dataset²⁶ released in April 2017. An MSA was generated for nine different bitscores, a sequence inclusion threshold was normalized to length and expressed as the number of bits per residue, with values ranging from the most inclusive (0.1) to the least inclusive (0.9). The alignment depth was chosen to optimize the number of non-redundant sequences with the fewest gaps in the

alignment as previously described^{9,10}, although the alignment choice was robust with respect to consistency of predicted evolutionary couplings over a wide range of alignment depths. Blindly optimizing the alignment choice resulted in two alignments, with the smaller one at 31,505 non-redundant sequences of length-346 residues (10–355) and an ‘effective number’ of 8,729 sequences after down-weighting sequences with more than 80% identity and no more than 30% gaps in any columns used for the EC model computation.

Model folding criteria. For the chosen MSAs, evolutionary couplings were determined using a pseudo-likelihood maximization (PLMC)^{27–29}. A mixture model approach identified the 99% percentile probability of being in contact. We used a combination of commonly used prediction methods to determine secondary structure predictions (PSIPRED³⁰ and PolyPhobius³¹), together with secondary structure propensity computed directly from local evolutionary couplings²⁹. This resulted in the identification of ten transmembrane helices and three smaller helices in ECL4, and two β-strands in ECL2. A total of 220 folded models for *T. thermophilus* RodA were generated for increasing numbers of evolutionary coupling restraints using the folding protocol in EVfold⁹, which uses a distance geometry and simulated annealing protocol in CNS^{32,33}. All models were ranked as previously described^{11,34} and the 50 top-ranked models from each of the MSAs were used as molecular replacement search models (see below). To generate additional models used for further ensemble-based phasing runs, five additional models were generated using the top-ranked fold prediction from bitscore 0.8 with an additional round of dedicated simulated annealing in CNS. Cartesian dynamics were used for heating, torsion for cooling and 600 final minimization steps with 50 cycles were used with other parameters kept as defaults. In addition to evolutionary couplings scores, for folded models EVfold generates an evolutionary coupling enrichment score for each residue that reflects how constrained that residue is, which can be thought of as conservation of ‘coupling’.

Phasing and refinement. One hundred evolutionary coupling-derived models of wild-type RodA were primed for molecular replacement using Sculptor³⁵, which included limited side-chain pruning and B-factor assignment based on accessible surface area³⁶. Each model was tested as a single search template for molecular replacement in Phaser¹², resulting in 98 out of the 100 initial models providing candidate solutions. The solutions were sorted by translation function Z-score (TF-Z) and the top 32 solutions manually inspected in PyMOL³⁷. A total of 22 solutions were highly similar and distinct from all other solutions. Sorting the solutions by log-likelihood gain (LLG) provided a similar result, in which all 22 solutions were among the top 40% of solutions. However, most of these were not well separated from other potential solutions in their respective searches, typically with two-to-four other candidate solutions giving LLG values at least 75% of that of the top solution. The models giving the top two solutions (as judged by both TF-Z and LLG metrics) were used as a single ensemble for phase calculation followed by manual building using COOT³⁸ and reciprocal space refinement using phenix.refine³⁹. Manual refinement was complemented by the use of ROSETTA refinement in Phenix¹³. Subsequently, additional cycles of manual building and reciprocal space refinement led to the final refined structure.

Verification of sequence register was straightforward and unambiguous owing to the relatively high resolution and frequency of bulky amino acid side chains (RodA is roughly 10% tryptophan, tyrosine and phenylalanine). Representative composite omit map density is shown in Extended Data Fig. 1. The structure of RodA(D255A) was solved using wild-type RodA as the molecular replacement search model and the resulting refined structure is nearly identical to that of wild-type RodA (0.1 Å r.m.s.d. between all C_α atoms). After refinement, the quality of both structures was assessed using MolProbity to calculate Ramachandran statistics and other parameters⁴⁰, and figures were prepared in PyMOL. All crystallographic data processing, refinement and analysis software was compiled and supported by the SBGrid Consortium⁴¹.

Predicting residue contacts between RodA and PBP2. EVComplex¹⁰ was used to predict inter-protein contacts between *T. thermophilus* full-length RodA and full-length PBP2 (Uniprot ID Q5SJ23). We constructed alignments for RodA (33,670 sequences) and PBP2 (40,764 sequences) as described above for RodA alone, using the April 2017 Uniprot release⁴² for clarity on species identifiers. We concatenated RodA and PBP2A sequences from each species when they were within 10,000 nucleotides of each other, based on European Nucleotide Archive data downloaded in February 2017⁴³, and evolutionary couplings were computed on the complex alignment as previously described¹⁰. A mixture model approach was used to identify top-scoring contacts, both within individual monomers and between RodA and PBP2, and to assign a probability to each contact of being within the tail of the distribution of coupling scores as previously described²⁹. This scoring method is accurate for predicting whether proteins interact as well as which residues are in contact¹⁰. The distribution of evolutionary couplings is approximated by a Gaussian log-normal mixture model, and we defined the tail of the distribution as those scores that have >95% probability of belonging to

the log-normal component. Evolutionary couplings in this tail defined as high probability resulted in 26 residue pairs predicted as contacts between RodA and PBP2.

Sequence and structure conservation analysis. The sequence conservation analysis shown in Fig. 3 and Extended Data Fig. 5 was computed using the ConSurf server⁴⁴. In brief, a multiple sequence alignment of *T. thermophilus* RodA to its closest 150 homologues was generated using the HHMER algorithm provided by ConSurf, with conservation scores plotted in PyMOL. Additionally, a multiple sequence alignment of RodA to 506 homologues from representative bacterial taxa was generated using a protein sequence BLAST search on the NCBI public database using *T. thermophilus* RodA protein sequence as query, and plotted in Extended Data Fig. 3.

Homology modelling and mutability index. A homology model of *B. subtilis* RodA based on the structure of *T. thermophilus* RodA was constructed using MODELLER⁴⁵. In brief, owing to the low sequence identity (~26%) between the two enzymes, a multiple sequence alignment of 10 RodA homologues from diverse bacterial taxa was performed (*T. thermophilus*, *B. subtilis*, *E. coli*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Shigella flexneri*, *Haemophilus influenzae*, *Deinococcus marmoris*, *Bacillus safensis* and *Lysinibacillus odyseyi*). The resulting alignment was used as the search template in MODELLER, in which disordered residues 204–256 and 266–276 were omitted from the model generation. A total of ten models were generated and the top-scoring model was chosen for further analysis.

The mutability index of each residue in the homology model of *B. subtilis* RodA was calculated based on previously reported MutSeq data¹. First, for each residue the nonsynonymous mutation rate was defined as the number of mutations observed after excluding synonymous and nonsense mutations. To calculate a raw mutability index, the nonsynonymous mutation rate was divided by the summed rate of silent mutations observed for a given nucleotide position. Finally, the raw mutability index for each residue was normalized with respect to the expected silent mutation rate for each general class of nucleotide change, resulting in a per cent mutability index for each residue. A low mutability index (that is, low percentage) represents a residue that is relatively intolerant of mutation at that position and a high mutability index (high percentage) represents a residue that is relatively tolerant of changes at that position.

RodA mutant plasmid and strain construction. Mutants of *E. coli* RodA were introduced using QuikChange mutagenesis into a plasmid containing the allele P_{lac} -pbpA-rodA(WT) (pHC857)^{1,2}. Wild-type *E. coli* cells (strain TB28) were then transformed with the mutant plasmids to generate strains that each harboured mutations in the central cavity (E114A, K117A, E114K/K117E and D159A) and at the RodA–PBP2 interface (L240S, S326A and L281A).

Mutations of *B. subtilis* RodA were introduced using QuikChange mutagenesis into the pER174a plasmid ($amyE::P_{xyl}$ -rodA(WT)-10×His)¹. The resulting mutagenic plasmids ($amyE::P_{xyl}$ -rodA(allele)-10×His) was directly transformed into the *B. subtilis* strain (rodA::kan, P_{spank} -rodA(WT)-10×His, and $P_{pensacA}::mCherry$) to generate strains each harbouring mutations in the central cavity (Q114A, E117A, K120A, E117K/K120E, D280N and S364A), the proposed lipid-binding cavities (I113A, I113S, and F118A and A171L), and the RodA–PBP2 interface (F292A). The *B. subtilis* strains harbouring $amyE::P_{xyl}$ -rodA-(D167A), $amyE::P_{xyl}$ -rodA(D167N), and $amyE::P_{xyl}$ -rodA(D280A) have previously been described¹.

Mutational analysis in vivo. All *E. coli* cultures were grown at 37°C in M9 minimal medium supplemented with 0.2% (w/v) maltose, 0.2% (w/v) casamino acids, and 25 µg ml⁻¹ chloramphenicol. Overnight cultures were diluted to an OD₆₀₀ of 0.05 and grown to OD₆₀₀ = 0.25 in the absence of inducer. These cultures were then further diluted to an OD₆₀₀ of 0.005 in medium containing 1 mM IPTG to induce expression of the pbpA-rodA construct. Cells were grown for five generations in the presence of inducer and fixed when the OD₆₀₀ reached between 0.15 and 0.2. Fixative solution contained 0.04% glutaraldehyde, 4% formaldehyde, 32 mM sodium phosphate, pH 7.5. Wide-field phase-contrast microscopy was performed on a Nikon TE2000 microscope equipped with a 100× Plan Apo 1.4 NA objective and a CoolSNAP HQ2 monochrome camera.

B. subtilis strains were derived from the PY79 prototrophic strain⁴⁶. Cells were grown in LB medium at 37°C in the presence of 10 µM IPTG to express wild-type RodA. When cultures reached mid-log, cells were washed three times with plain LB medium and resuspended in 20 ml of LB medium to an OD₆₀₀ of 0.02, then supplemented with 10 mM xylose to induce expression of RodA wild type and variants. Cells were analysed by phase-contrast microscopy 90 min later. Microscopy was performed on a Nikon Ti microscope equipped with Plan Apo 100×/1.4NA phase-contrast oil objective and a CoolSNAP HQ2 camera. Cells were immobilized using 2% agarose pads, containing growth medium. Images were cropped and adjusted using MetaMorph software (Molecular Devices).

Mutant analysis in vitro. Mutants of *B. subtilis* RodA were transformed into *E. coli* strain CAM333, a derivative of strain C43 with deletions in *ponB*, *pbpC* and

mgtA, and were expressed and purified by immunoaffinity chromatography as previously described¹.

Before circular dichroism spectroscopy experiments, each RodA mutant was dialysed into a buffer consisting of 10 mM sodium phosphate pH 7.5, 500 mM potassium fluoride, 0.5% (w/v) CHAPS, and 0.05% (w/v) DDM. Spectra were acquired on a Jasco J-815 spectropolarimeter. Circular dichroic spectra were recorded between 200–260 nm using a quartz cuvette with a path length of 1 mm, a 50-nm/min scanning speed, and a bandwidth of 1 nm. Five spectra were measured at 25°C, averaged and corrected for buffer contribution. Secondary structure assessment was not performed owing to high absorbance contributions at wavelengths less than 205 nm from the detergent mixture necessary for protein stability and function (0.5% CHAPS, 0.05% DDM).

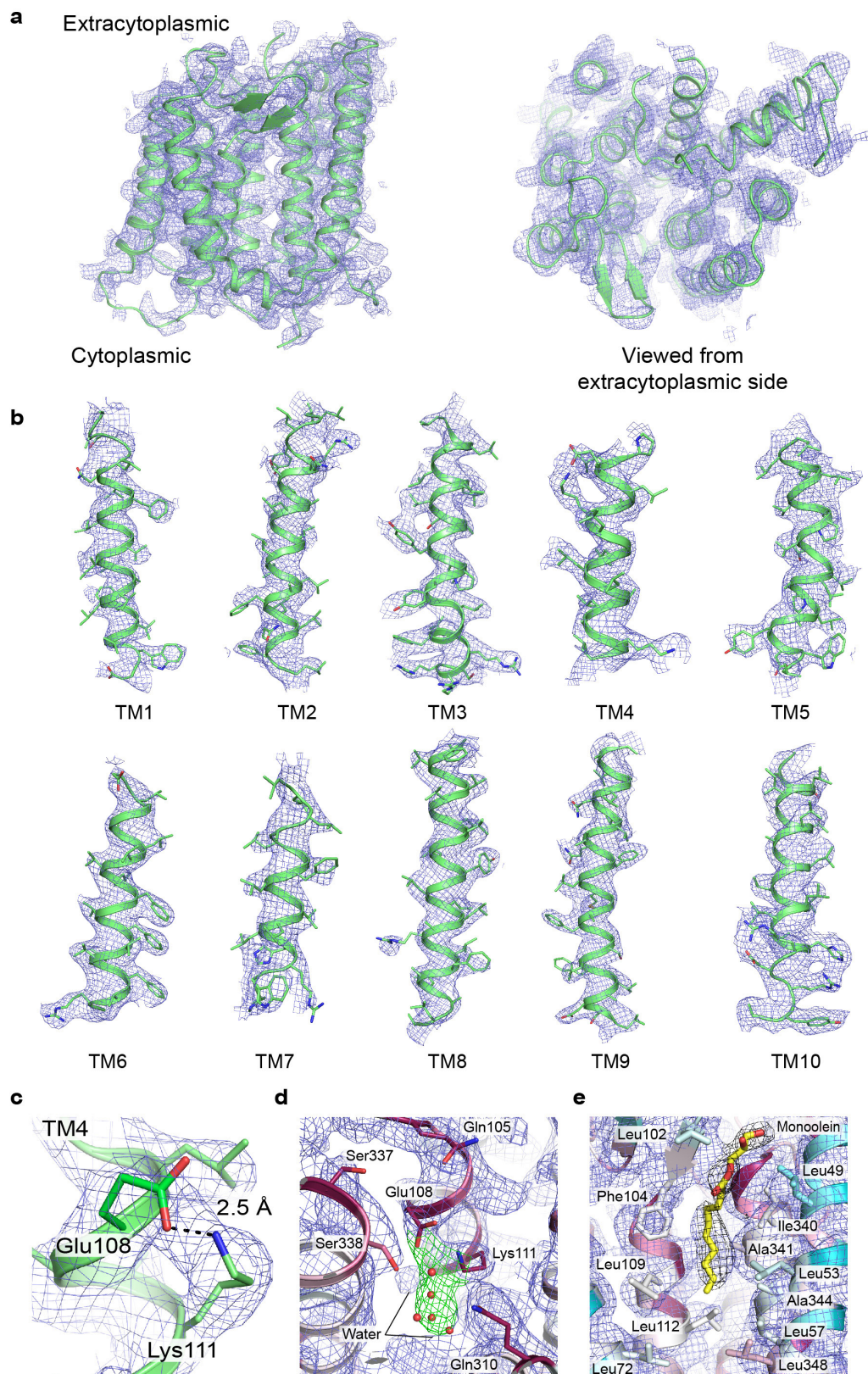
For assessment of enzymatic activity, lipid II substrate was purified from *E. coli* as described⁴⁷. Peptidoglycan polymerization reactions were adapted from previously described methods⁴⁸. In brief, purified *B. subtilis* wild-type and mutant RodA proteins were incubated with lipid II in reaction buffer containing 50 mM HEPES pH 7.0, 20 mM CaCl₂, 20 mM MgCl₂ and 20% DMSO. The working concentration of RodA and lipid II were 1 µM and 20 µM, respectively. RodA was purified in 0.5% CHAPS and 0.05% DDM, and therefore the working concentration for CHAPS and DDM was 0.05% and 0.005%, respectively. As a positive control lipid II was also polymerized with SgtB(Y181D) (1 µM) in reaction buffer containing 12.5 mM HEPES pH 7.5, 2 mM MnCl₂, 0.25 mM Tween-80 and 20% DMSO. All reactions were incubated at 25°C for 1 h and quenched by incubation at 95°C for 2 min. Peptidoglycan biotinylation of each reaction mixture was performed by addition of biotinylated D-lysine (1.5 mM, final concentration) and PBP4 (3.8 µM, final concentration) followed by incubation at 25°C for 1 h. The biotinylation reaction was then quenched by addition of 13 µl 2× SDS loading dye. The samples were then loaded into a 4–20% gradient polyacrylamide gel and run at 180 V. The products were transferred onto a PVDF membrane (BioRad) and fixed in 0.4% paraformaldehyde diluted in PBS for 30 min at room temperature. The membrane was blocked with SuperBlock TBS blocking buffer (Thermo Fisher Scientific) for 1 h at room temperature and the biotinylated products were detected by incubation with fluorescently tagged streptavidin (IRDye 800-CW streptavidin (Li-Cor Biosciences, 1:5,000 in SuperBlock)) for an additional 30 min. Membranes were washed 4 × 10 min with TBST (0.01% Tween 20) and given a final 10-min wash in PBS. Blots were then visualized using an Odyssey CLx imaging system (LI-COR Biosciences).

Code availability. The full EVfold software package is available at <https://github.com/debbiemarkslab/EVcouplings>.

Data availability. Structure factors and refined atomic coordinates for RodA wild type and the RodA(D255A) mutant are deposited in the RCSB Protein Data Bank under accession codes 6BAR and 6BAS, respectively.

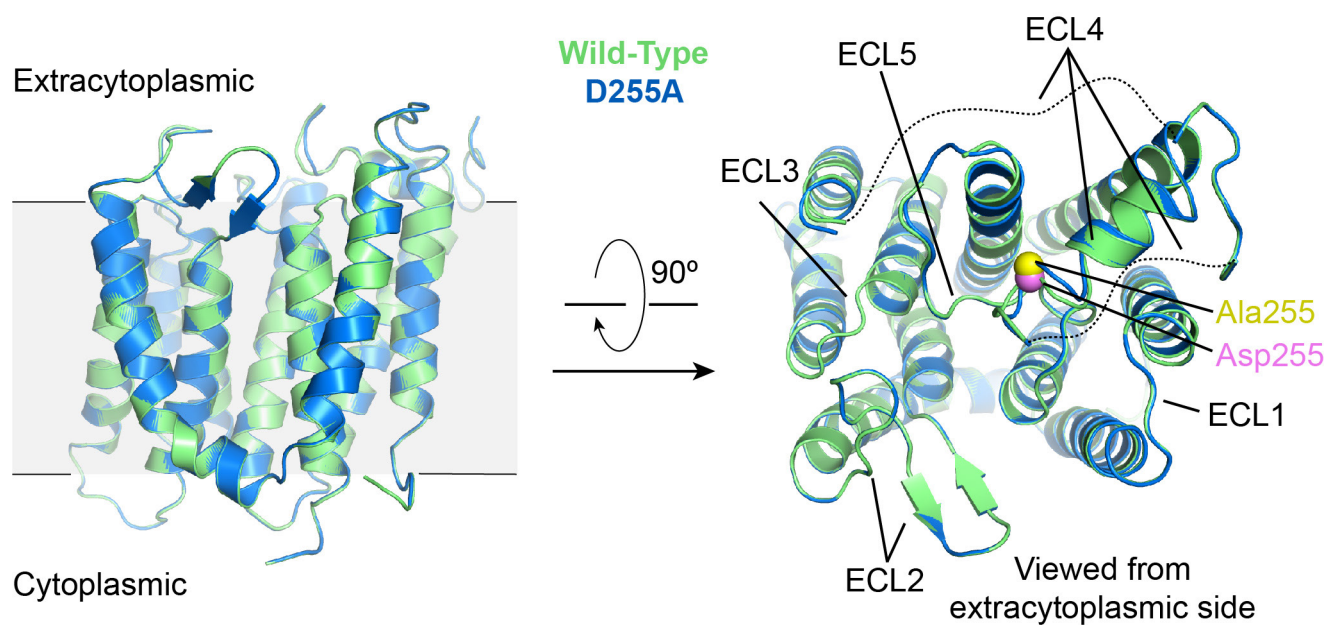
23. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nat. Protoc.* **4**, 706–731 (2009).
24. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
25. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
26. Supek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
27. Ekeberg, M., Lökvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
28. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
29. Toth-Petroczy, A. et al. Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.e12 (2016).
30. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
31. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432 (2007).
32. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2**, 2728–2733 (2007).
33. Brünger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
34. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
35. Bunkóczi, G. & Read, R. J. Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr. D* **67**, 303–312 (2011).
36. Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. The importance of alignment accuracy for molecular replacement. *Acta Crystallogr. D* **60**, 1229–1236 (2004).
37. Schrödinger. *The PyMOL Molecular Graphics System, Version 1.3r1* (Schrödinger, 2010).

38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
39. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
40. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
41. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
42. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
43. Pakseresht, N. *et al.* Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res.* **42**, D38–D43 (2014).
44. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529–W533 (2010).
45. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **54**, 5.6.1–5.6.37 (2016).
46. Youngman, P. J., Perkins, J. B. & Losick, R. Genetic transposition and insertional mutagenesis in *Bacillus subtilis* with *Streptococcus faecalis* transposon Tn917. *Proc. Natl Acad. Sci. USA* **80**, 2305–2309 (1983).
47. Qiao, Y. *et al.* Lipid II overproduction allows direct assay of transpeptidase inhibition by β -lactams. *Nat. Chem. Biol.* **13**, 793–798 (2017).
48. Qiao, Y. *et al.* Detection of lipid-linked peptidoglycan precursors by exploiting an unexpected transpeptidase reaction. *J. Am. Chem. Soc.* **136**, 14678–14681 (2014).



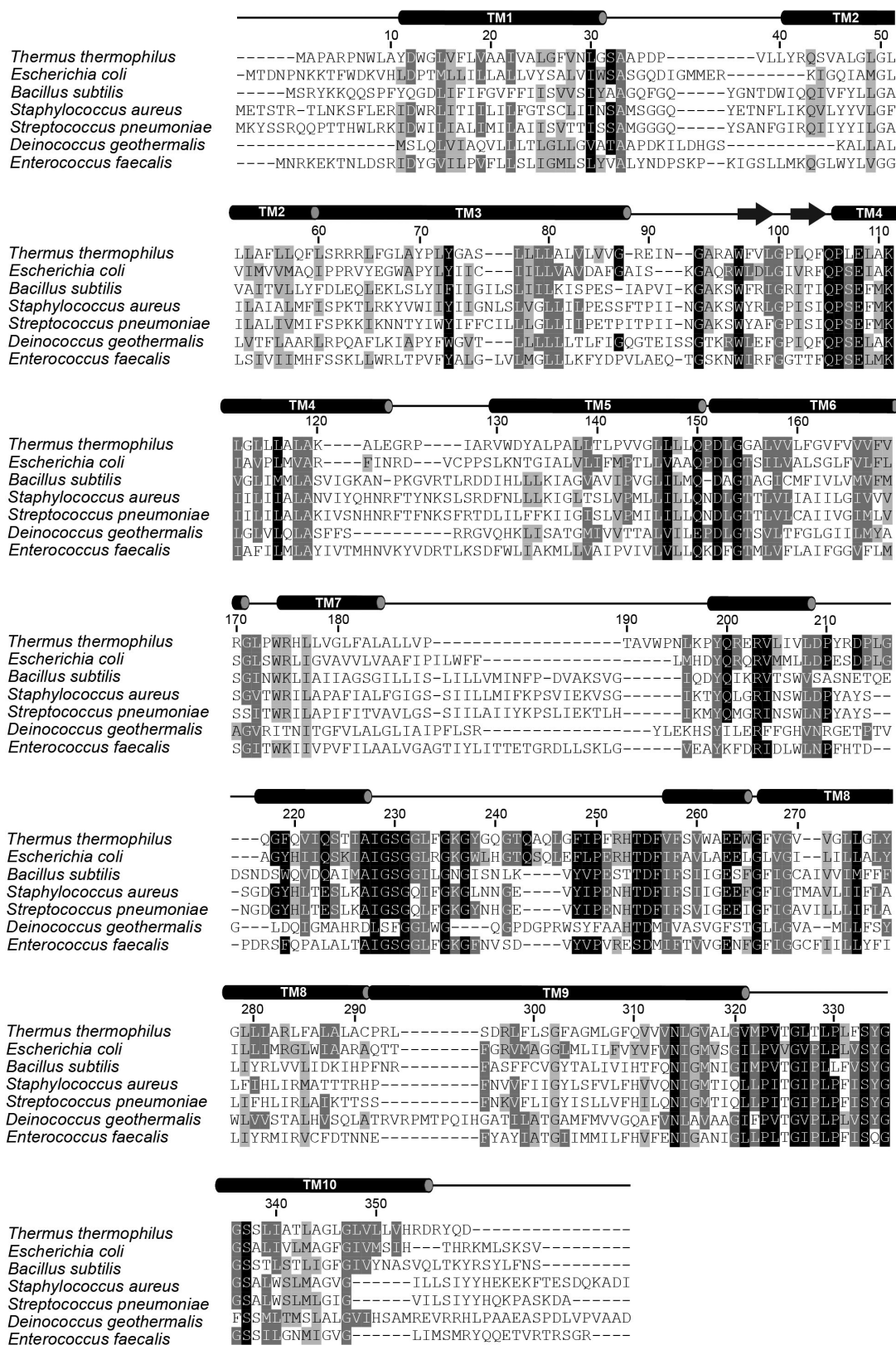
Extended Data Figure 1 | Representative electron density. **a–c**, Simulated annealing composite omit $2F_o - F_c$ electron density map of *T. thermophilus* RodA contoured at 1.0σ within a 2.0 \AA radius of atoms shown. **d**, The same map contoured at 1.0σ and coloured blue and green for RodA and

water molecules, respectively. The modelled water molecules are shown as red spheres. **e**, The same map contoured at 1.0σ within a 3.0 \AA radius RodA (shown in blue) and the same map contoured at 1.0σ for monoolein (shown in black).



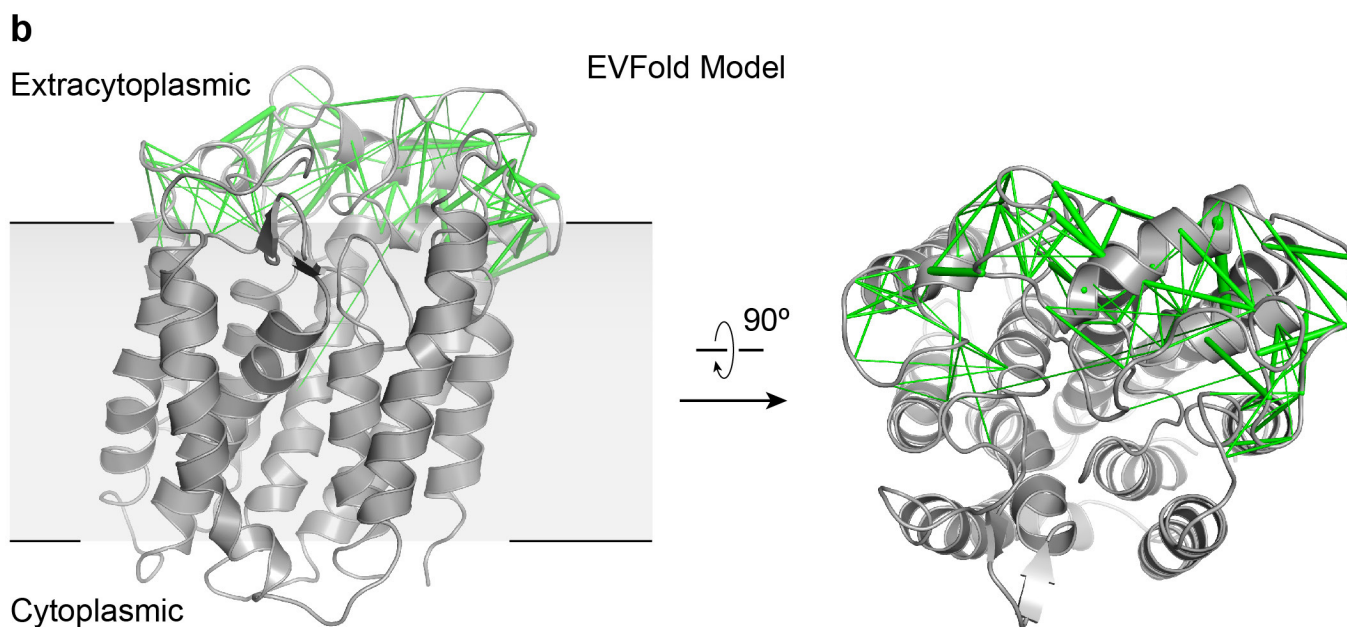
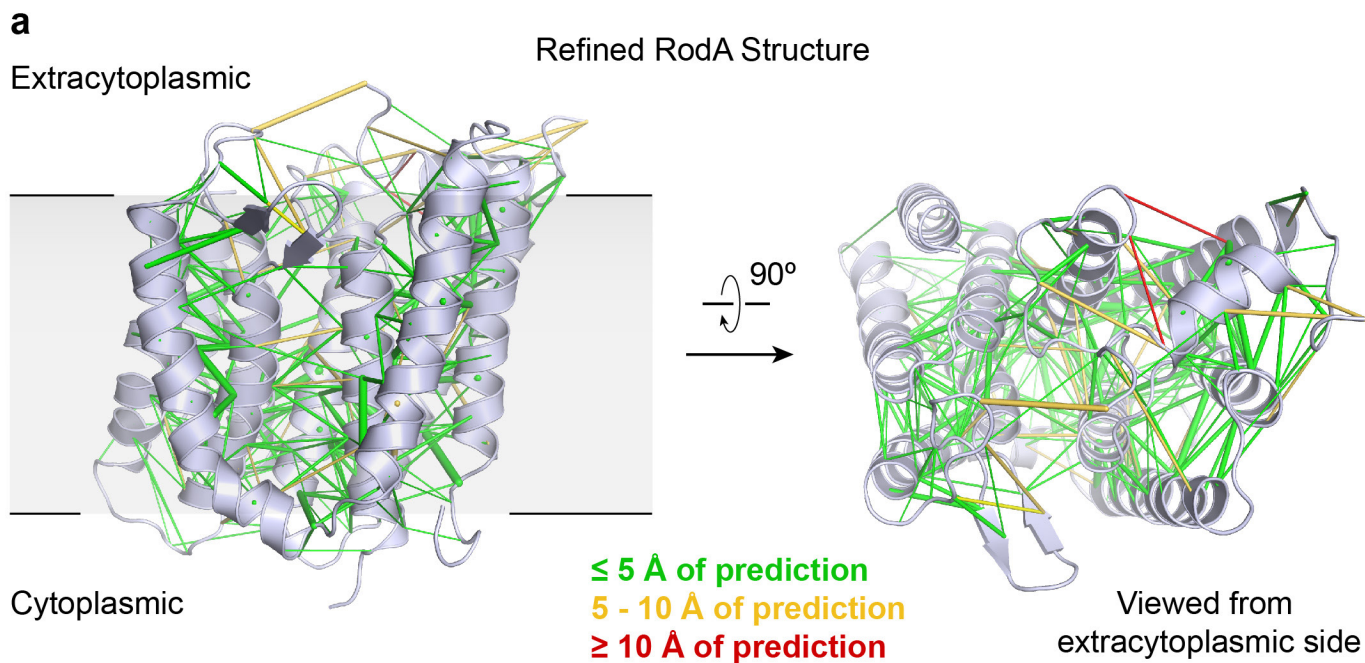
Extended Data Figure 2 | Comparison of RodA wild type and RodA(D255A) structures. Structures of wild-type (green) and D255A (blue) RodA are shown viewed parallel to the membrane (left) and from the extracytoplasmic side (right). The C_{α} atom of residue 255 for each

structure is shown as a sphere, and coloured pink (wild type) or yellow (D255A). The dashed lines represent the disordered residues 189–227 and 237–251 in both structures. The two structures are essentially identical, with a C_{α} r.m.s.d. of 0.1 Å.



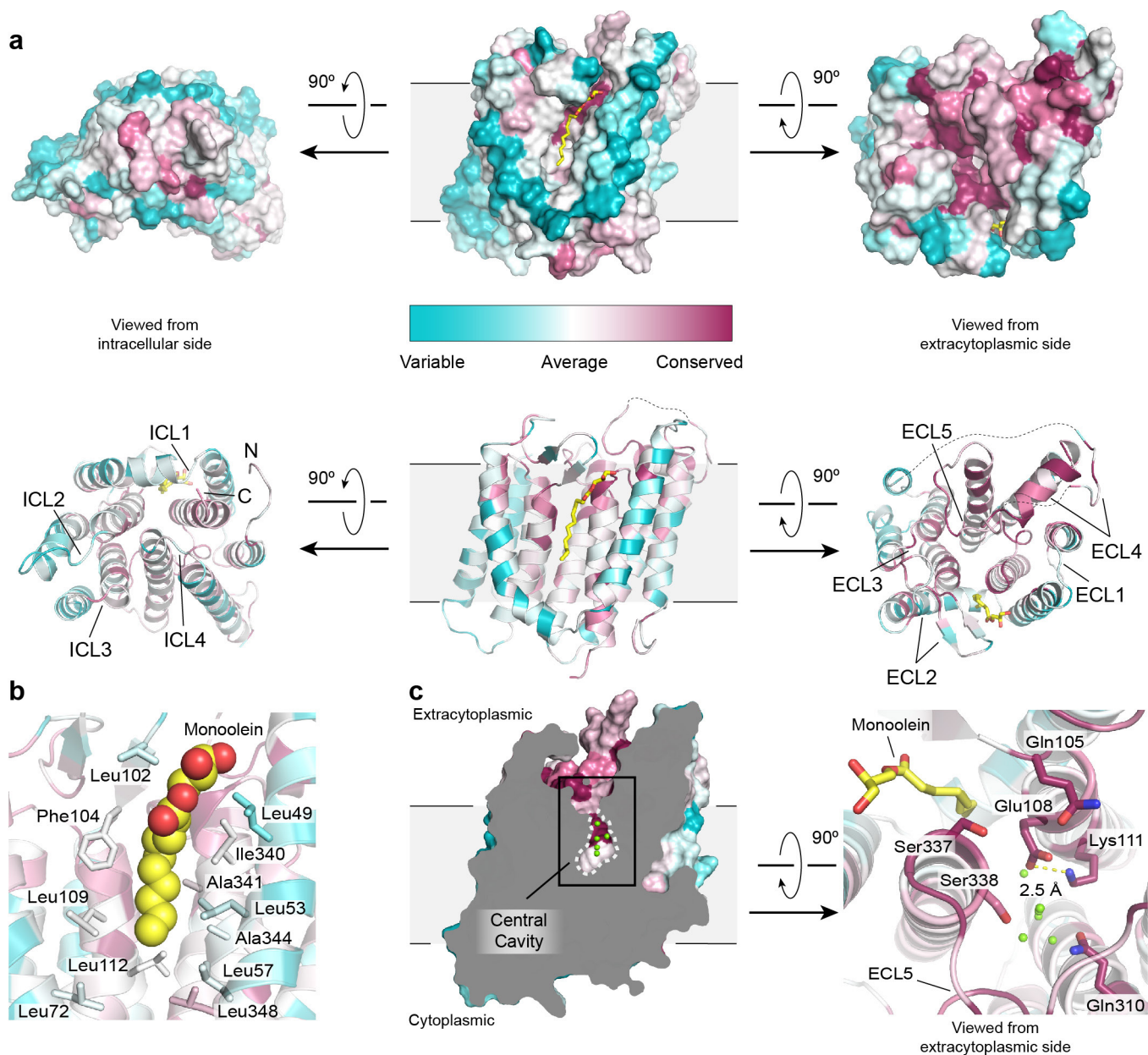
Extended Data Figure 3 | RodA sequence conservation. The results of an alignment of 506 RodA sequences from diverse bacterial taxa, with representative examples displayed. Residues with 98%, 80% and 60% similarity across all 506 sequences are shown in black, grey and light grey, respectively.

Secondary structure elements are shown above the alignment on the basis of the *T. thermophilus* RodA crystal structure and JPRED analysis of the portions of ECL4 that were not modelled in the structure.



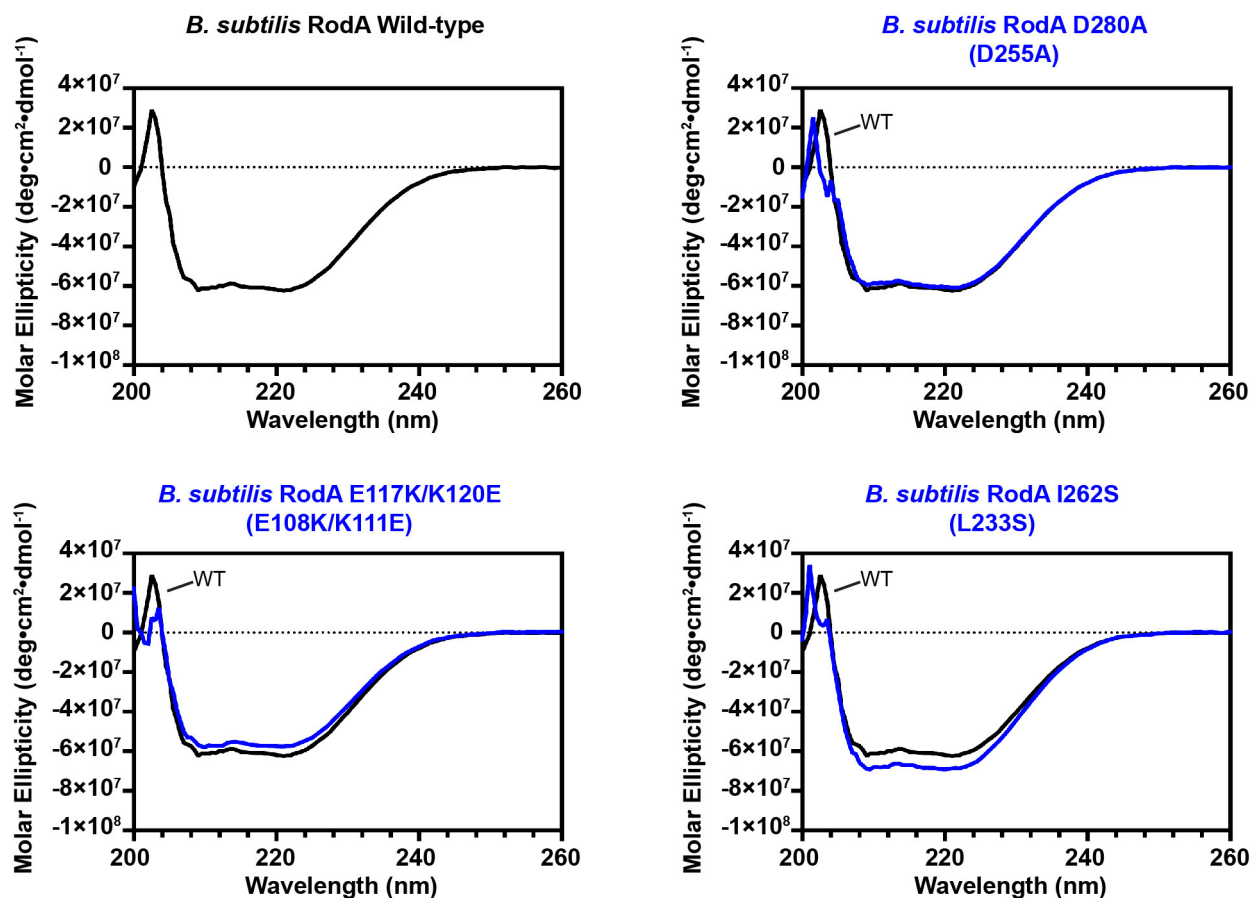
Extended Data Figure 4 | RodA evolutionary couplings. a, The refined crystal structure of *T. thermophilus* RodA (shown in light blue) is in close agreement with its evolutionary couplings. Green, yellow and red lines between residues represent regions of the structure that are less than 5 Å, between 5 and 10 Å, and greater than 10 Å of the predicted evolutionary

couplings, respectively. **b,** The partially disordered ECL4 in the refined structure is strongly coupled evolutionarily to the transmembrane domain of RodA. A representation of predicted intra- and inter-domain evolutionary couplings for ECL4 is mapped onto a EVfold model (shown in grey).



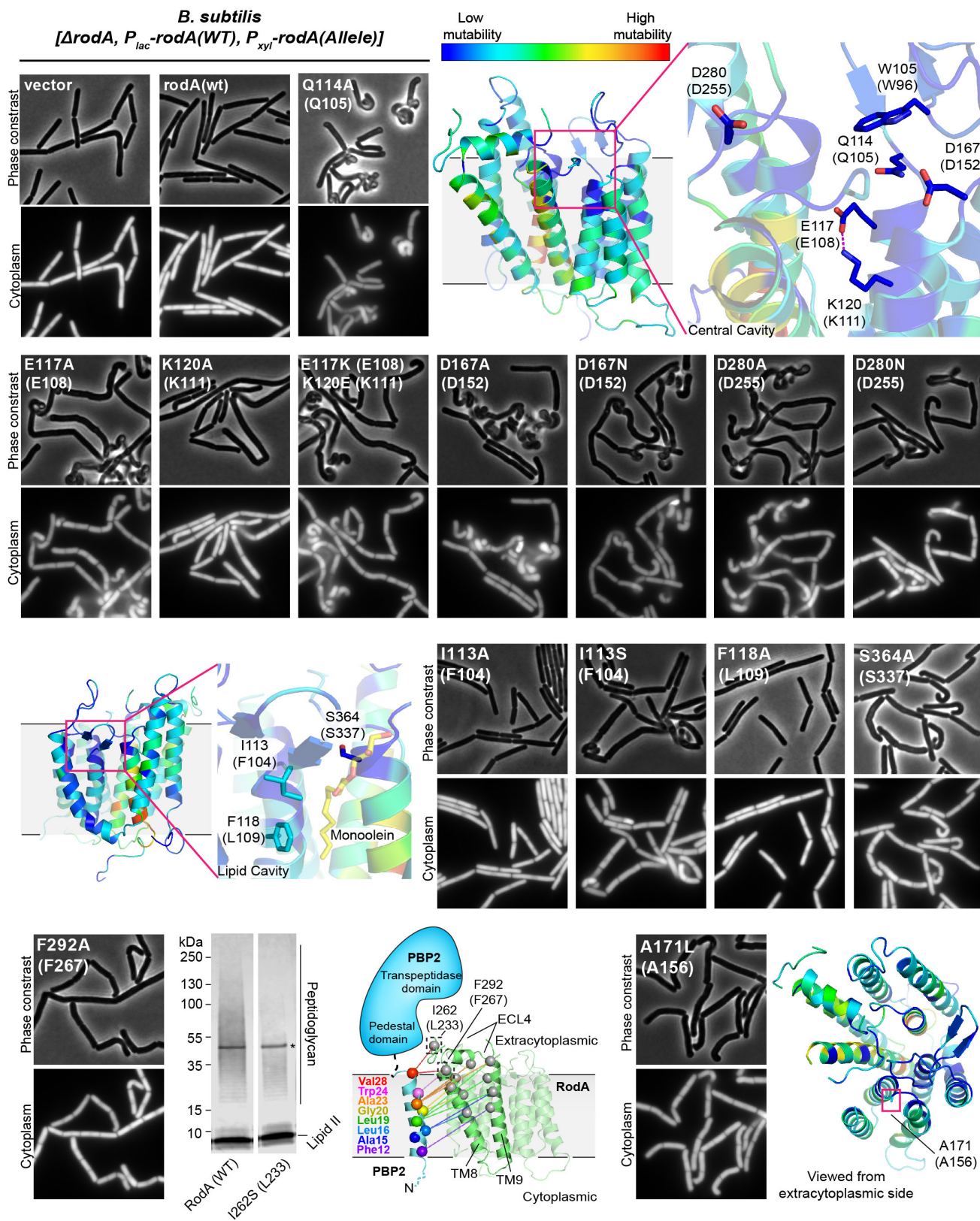
Extended Data Figure 5 | Sequence conservation of *T. thermophilus* RodA. a, Surface and ribbon representation of RodA (top and bottom, respectively). Analysis was performed using Consurf, coloured in a scale from teal (poorly conserved) to magenta (highly conserved). A bound lipid modelled as monoolein is shown in yellow sticks. The dashed lines

represent disordered residues 189–227 and 237–251 in ECL4. **b**, The bound lipid (shown as spheres) is surrounded by many aliphatic amino acids (shown as sticks). **c**, Extracytoplasmic view of the water-filled central cavity and its proximity to the bound lipid molecule.



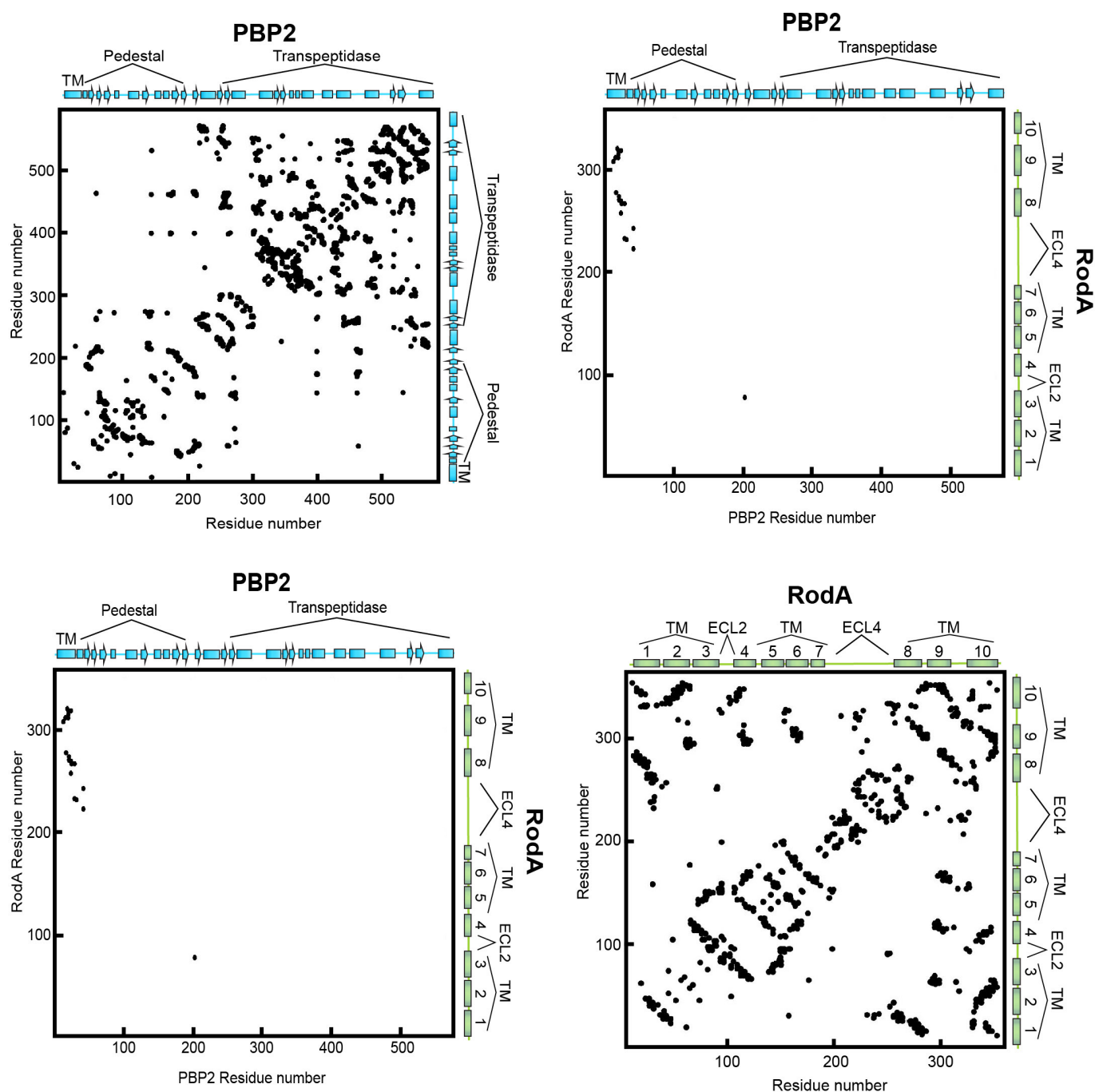
Extended Data Figure 6 | Circular dichroism spectroscopy analysis of purified *B. subtilis* RodA variants. Circular dichroism measurements of wild-type RodA as well as RodA(D280A), RodA(E117K/K120E) and RodA(I262S) indicate that the overall folds of all four forms are comparable and display characteristic α -helical peaks at 208 and 222 nm.

For each panel, the circular dichroic spectra of wild-type and the indicated mutant RodA are shown in black and blue, respectively. The corresponding *T. thermophilus* numbering for each mutant is shown in parentheses. Experiments were repeated independently twice with similar results.



Extended Data Figure 7 | Mutagenic analysis of *B. subtilis* RodA function *in vivo*. Micrographs of *B. subtilis* strains harbouring an IPTG-inducible allele of wild-type *rodA*, and wild-type or mutant P_{xyIA}-*rodA*. Expression was induced by growing cells in the presence of 10 μ M IPTG and 10 mM xylose. The bacterial cytosol is indicated by intracellular mCherry expression (P_{pen}-mCherry). Mutants in the central cavity show

particularly deleterious phenotypes in this dominant negative assay. Experiments were repeated independently 2–3 times with similar results. A mutation made in the predicted RodA–bBPP interface does not prevent peptidoglycan polymerization *in vitro* (lower left panel), representative of two independent experiments.



Extended Data Figure 8 | Evolutionary coupling analysis for the RodA-PBP2 complex. Evolutionary co-variation maps highlighting 693 couplings for PBP2 (top left panel), and 362 couplings for RodA (bottom right panel) using a 95% confidence threshold. These maps display good

correlation to the crystal structure of RodA and homology models of PBP2. The top right and lower bottom panels depict the 19 predicted inter-protein contacts between RodA and PBP2 using the same 95% confidence threshold.

Extended Data Table 1 | Data collection and refinement statistics

	<i>T. thermophilus</i> RodA Wild-type	<i>T. thermophilus</i> RodA D255A
Data collection		
Space group	C2	C2
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	122.4, 80.0, 47.8	121.2, 79.2, 47.4
α , β , γ (°)	90.0, 91.1, 90.0	90.0, 91.1, 90.0
Resolution (Å)	40.0 - 2.9 (3.1 - 2.9)*	40.0 - 3.2 (3.4 - 3.2)*
R_{sym} (%)	10.6 (147.9)	17.1 (204.9)
$\langle I/\sigma(I) \rangle$	6.69 (0.69)	4.08 (0.44)
$CC_{1/2}$ (%)	99.7 (45.8)	99.6 (41.7)
Completeness (%)	98.4 (98.8)	99.6 (99.5)
Multiplicity	3.02 (3.15)	3.43 (3.35)
Refinement		
Resolution (Å)	40.0 - 2.91 (2.99 - 2.91)	40.0 - 3.19 (3.31 - 3.19)
No. reflections	18684 (1851 in test set)	13756 (1360 in test set)
$R_{\text{work}} / R_{\text{free}}$ (%)	22.9 / 27.4	27.7 / 30.2
No. atoms		
Protein	2229	2201
Solvent ions/lipids	44	1
Water	52	8
Average <i>B</i> -factors (Å ²)		
Protein	104.2	111.0
Lipids	109.6	-
Solvent ions	105.4	125.9
Waters	87.6	89.9
R.m.s. deviations		
Bond lengths (Å)	0.005	0.002
Bond angles (°)	0.825	0.751
Ramachandran Statistics		
Favored (%)	94.5	93.5
Allowed (%)	5.5	6.5
Outliers (%)	0.0	0.0

Each dataset was collected from a single crystal.

*Values in parentheses are for the highest-resolution shell.

Structure of the insulin receptor–insulin complex by single-particle cryo-EM analysis

Giovanna Scapin^{1*}, Venkata P. Dandey^{2*}, Zhening Zhang², Winifred Prossise¹, Alan Hruza¹, Theresa Kelly³, Todd Mayhood³, Corey Strickland¹, Clinton S. Potter² & Bridget Carragher²

The insulin receptor is a dimeric protein that has a crucial role in controlling glucose homeostasis, regulating lipid, protein and carbohydrate metabolism, and modulating brain neurotransmitter levels^{1,2}. Insulin receptor dysfunction has been associated with many diseases, including diabetes, cancer and Alzheimer's disease^{1,3,4}. The primary sequence of the receptor has been known since the 1980s⁵, and is composed of an extracellular portion (the ectodomain, ECD), a single transmembrane helix and an intracellular tyrosine kinase domain. Binding of insulin to the dimeric ECD triggers autophosphorylation of the tyrosine kinase domain and subsequent activation of downstream signalling molecules. Biochemical and mutagenesis data have identified two putative insulin-binding sites, S1 and S2⁶. The structures of insulin bound to an ECD fragment containing S1 and of the apo ectodomain have previously been reported^{7,8}, but details of insulin binding to the full receptor and the signal propagation mechanism are still not understood. Here we report single-particle cryo-electron microscopy reconstructions of the 1:2 (4.3 Å) and 1:1 (7.4 Å) complexes of the insulin receptor ECD dimer with insulin. The symmetrical 4.3 Å structure shows two insulin molecules per dimer, each bound between the leucine-rich subdomain L1 of one monomer and the first fibronectin-like domain (FnIII-1) of the other monomer, and making extensive interactions with the α -subunit C-terminal helix (α -CT helix). The 7.4 Å structure has only one similarly bound insulin per receptor dimer. The structures confirm the binding interactions at S1 and define the full S2 binding site. These insulin receptor states suggest that recruitment of the α -CT helix upon binding of the first insulin changes the relative subdomain orientations and triggers downstream signal propagation.

The insulin receptor is a dimer of heterodimers that comprises two α -chains and two β -chains⁵, represented as $(\alpha\beta)_2$. The α -chain and approximately 190 residues at the N-terminus of the β -chain are located on the extracellular side of the plasma membrane (Fig. 1a) and together constitute the full insulin receptor ECD. The remainder of the β -chain includes a transmembrane helix, the juxtamembrane domain and the intracellular tyrosine kinase domain. Structures of the individual subdomains that make up the ECD (leucine-rich (L)1 and L2, cysteine-rich (CR) and fibronectin type III domains FnIII-1, FnIII-2 and FnIII-3) have been resolved, and the crystal structure of the apo ECD⁸ provides one view of their quaternary organization. In the apo ECD structure, density for the insertion domain (120 residues located within FnIII-2, containing the cleavage site that generates the α - and β -chains and the α -CT helix) was poor, and only a portion of the α -CT helix was visible^{9,10}. As only one $(\alpha\beta)$ heterodimer is present in the asymmetric unit of the insulin receptor crystal structure, a symmetry-generated tetramer, $(\alpha\beta)_2$, has been hypothesized as representing the biologically relevant form of the ECD⁸. Hereafter, we refer to each individual

$(\alpha\beta)$ heterodimer as the insulin receptor monomer, and to the $(\alpha\beta)_2$ tetramer as the insulin receptor dimer.

Current models of insulin binding to the insulin receptor suggest that there are two interaction sites on insulin that engage with two binding sites (S1 and S2) on the receptor¹¹. Binding is thought to occur *in trans*, with each insulin molecule interacting with S1 of one insulin receptor monomer and S2 of the other monomer. The bridging of these receptor binding sites constitutes the high-affinity interactions, whereas low-affinity interactions arise from single-site occupancy. The high-affinity interaction has a K_d of 6–200 pM for the solubilized and affinity-purified full-length receptor, whereas the low-affinity interactions have K_d values of 6 nM (for S1) and 400 nM (for S2)¹². Upon insulin binding to the full-length receptor, there are multiple binding events and dissociations and/or multiple rearrangements within low- and high-affinity binding poses¹². Receptor activation is caused by high-affinity binding; however, the activation curve is complex and shows anti-cooperative binding¹³.

Here we report the 4.3 Å resolution structure of the insulin receptor ECD in the presence of insulin, obtained using single-particle cryo-electron microscopy (cryo-EM) (Fig. 1b, c). A pronounced preferred orientation of the complex in vitreous ice was overcome with a combination of tilted data collection¹⁴ and fast plunge speeds using the Spotiton instrument^{15,16}. Further details are provided in the Methods. Three-dimensional classification of the cryo-EM data identified a dimeric structure with two populations, in an approximately 4:1 ratio (Fig. 2, Extended Data Fig. 1). The larger of these populations, class 1, is characterized by a symmetrical 'head' and a poorly resolved, apparently flexible 'stalk'. Refinement of the class 1 structure, after masking out flexible portions of the stalk and applying C2 symmetry, provided a reconstruction to 4.3 Å (Extended Data Figs 1, 2). The electron density map enabled unambiguous identification of the L1, CR, L2 and FnIII-1 receptor domains, as well as the clear positioning of the insulin molecule and the α -CT helix (Fig. 1b, c). Refinement of the data using C1 symmetry produced a lower resolution map (4.7 Å; Extended Data Fig. 3) in which additional density, assigned to FnIII-2, was visible for one monomer (Extended Data Figs 4a, 5b).

The class 1 structure shows two insulin molecules, symmetrically bound, per dimeric receptor. Although asymmetric binding was predicted for the full-length receptor¹², studies show that the soluble ECD binds two molecules of insulin, each with a K_d of 3.5 nM and a fast dissociation rate¹⁷. The symmetrical binding of both insulin molecules observed in the cryo-EM map is consistent with these data. Each insulin molecule is located between the L1 domain of one receptor monomer, the FnIII-1 domain of the other monomer, and the α -CT helix (Fig. 1c). Residues of the L1 subdomain and the α -CT helix, previously identified biochemically as essential for insulin binding¹³, have been hypothesized to represent the S1 site¹⁸. Structures of 'S1 microreceptors' (containing

¹Merck & Co., Department of Biochemical Engineering & Structure, 2000 Galloping Hill Road, Kenilworth, New Jersey 07033, USA. ²Simons Electron Microscopy Center, National Resource for Automated Molecular Microscopy, New York Structural Biology Center, 89 Convent Avenue, New York, New York 10027, USA. ³Merck & Co., Department of Biophysics, NMR & Protein Products Characterization, 2000 Galloping Hill Road, Kenilworth, New Jersey 07033, USA.

*These authors contributed equally to this work.

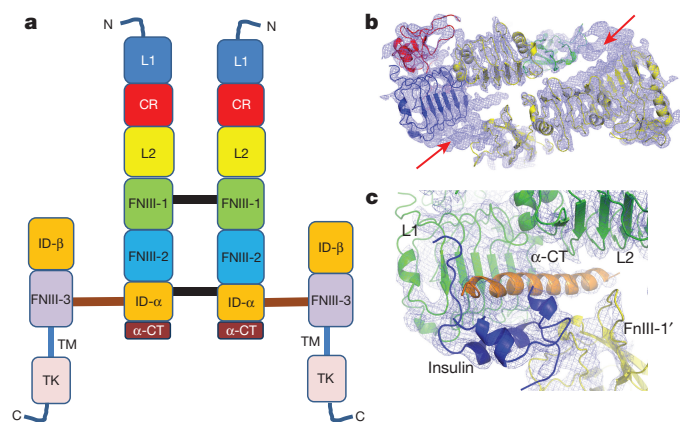


Figure 1 | Structure of the insulin receptor dimer. **a**, Domain organization of the full-length insulin receptor. The inter-monomer disulfide bonds^{28,29} are shown as black lines; the intra-monomer 'signalling bridge'²⁶ is shown as an orange line. TM, transmembrane domain; TK, tyrosine kinase domain. **b**, Cryo-EM density map for class 1 with the insulin receptor subdomains fitted to the density; one monomer is shown in yellow, the other is colour coded as in **a**. Density indicated by the red arrows can be attributed to the insulin and α -CT helix. **c**, Close-up of density with insulin and α -CT helix fitted.

L1, CR and a portion of the α -CT) in complex with insulin or insulin mimetics⁷ provided the first atomic details of some key insulin receptor–insulin interactions (Extended Data Table 1). In these structures, insulin interacts directly with the L1 subdomain of the receptor through only a small region of the B helix, and most of the other residues in the receptor that have been identified as essential for insulin binding (Extended Data Table 1) interact with the α -CT helix. Predicted interactions involving the C-terminal end of the insulin β -chain⁶ were also structurally visualized (Extended Data Table 1). The cryo-EM structure confirms the engagement of S1, including the interactions involving the C terminus of the insulin β -chain, which is now resolved. The C-terminal region of the β -chain in the bound form assumes a different conformation from that observed in the insulin-free form (for example, Protein Data Base (PDB) ID: 1ZNI; Extended Data Fig. 4c), allowing the β -chain core residues Gly8–Cys19 to interact with the α -CT helix, confirming the so-called detachment model of insulin binding¹⁹. Although it is not possible to determine which receptor monomer each α -CT helix originates from using the current map, the fact that each insulin molecule interacts with residues belonging to both monomers supports the *trans*-binding-mode hypothesis.

Analysis of the cryo-EM structure shows that S2 (Extended Data Table 1) is defined by interactions between the insulin receptor sequences Pro495–Arg498 and Arg539–Asn541 (located within the FnIII-1 domain) and chain B of insulin (residues Gln4–Gly8 and His10). Earlier work suggested that these insulin residues interacted with the S2 binding site⁶, but no partner residues were assigned on the receptor. The present structure shows interactions between the α -CT helix residues Leu696–Lys703 and insulin receptor residues Gly346–Asn349 and Arg372–Tyr374 (in the L2 subdomain) and Arg498–Leu501 and Val570–Thr571 (in the FnIII-1 subdomain). Previous work had correctly predicted that residues in L2 and FnIII-1 make up the S2 site²⁰. Clinical mutations (S323L, F382V, K460E and N462S, all located within or near the L2 and FnIII-1 subdomains discussed above) result in impaired insulin binding or reduced signalling^{21–24}. Further evidence that these regions represent S2 comes from ref. 25, in which a truncated receptor formed from the L1, CR, L2 and FnIII-1 subdomains fused to the α -CT helix was shown to have insulin-binding properties that are identical to those of the full ECD. A second possible location for the S2 site (at the junction of the FnIII-1 and FnIII-2 domains) was hypothesized from the crystallographic structure of the insulin-free insulin receptor ECD dimer⁸. Mutagenesis data provided some support for this assignment¹¹; however, in the un-symmetrized map (Extended

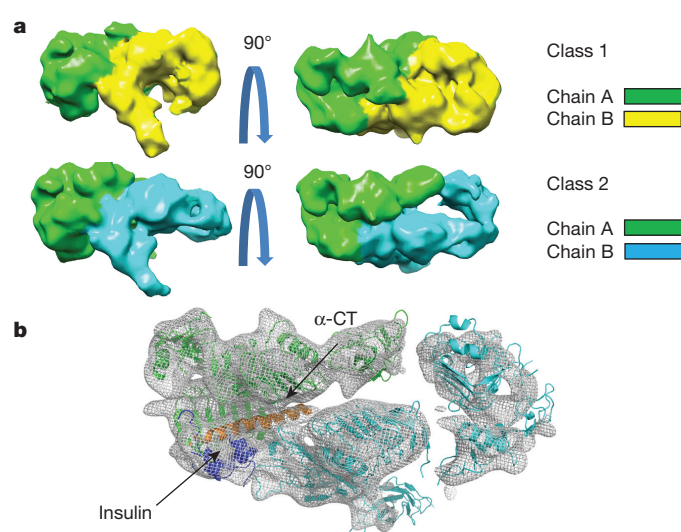


Figure 2 | Class 1 and class 2 dimers. **a**, Side and top views of the two classes identified during 3D reconstruction. **b**, Cryo-EM map for class 2, showing density for the α -CT helix and insulin only in one monomer.

Data Figs 4a, 5b) in which the FnIII-2 is visible, the bound insulin is more than 50 Å from this predicted S2 site (Extended Data Fig. 4b), on the opposite side of the receptor.

Refinement of the structural data for the smaller class 2 produced a 7.4 Å map (Fig. 2, Extended Data Figs 1, 6). In this map, half of the 'head' has an overall conformation similar to that observed in the class 1 map, with visible densities for the α -CT helix and insulin. The other half showed the L1–CR–L2 portion in a different conformation (Fig. 2a): more open, and characterized by less-well-defined density. The model of the insulin receptor based on the class 1 structure was rigid-body fitted onto the class 2 map, and only one insulin molecule and one α -CT helix were included in the final structure. The open side lacked both the insulin and the helix (Fig. 2b). This conformation therefore comprises both an insulin-bound receptor monomer and an insulin-free receptor monomer. Comparison of these structures suggests that insulin binding induces both the closing of the top portion of the receptor (by rigid-body motion of the L1–CR–L2 portion with respect to the FnIII-1 domain) and recruitment of the α -CT helix, contrary to a previously suggested mechanism in which insulin docks to a preformed 'harbour' containing both the L1 and α -CT elements that are required for binding²⁶.

The insulin receptor dimer identified in the cryo-EM analysis does not resemble the crystallographic symmetry-generated dimer⁸. The relative arrangement of the two monomers differs between the cryo-EM and crystallographic dimers, and transitioning between them would require major conformational changes and disruption of extensive surface interfaces, specifically the interactions between L1 and FnIII-2' and their dimeric symmetry mates, and between L2, FnIII-1' and their dimeric symmetry mates (Fig. 3a and Extended Data Fig. 7). However, the overall conformation of the crystallographic monomer (PDB ID: 4ZXB) is similar to that of the unbound monomer observed in the class 2 cryo-EM structure (Fig. 3b), suggesting that it could be a biologically relevant representation of the unbound insulin receptor monomer. Attempts to obtain a cryo-EM reconstruction for the ECD in the absence of insulin (the apo state) proved unsuccessful as the particles showed a high degree of heterogeneity and many were smaller than expected, suggesting that the ECD dimer may be unstable in the absence of the ligand. Comparing the unbound monomer from the crystal structure to the insulin-bound monomer from the cryo-EM structure (Extended Data Fig. 8), it appears that the conformational change that occurs upon insulin binding can be described as two rotations. The first is a 35° rotation of the L2–CR–L1 domains with

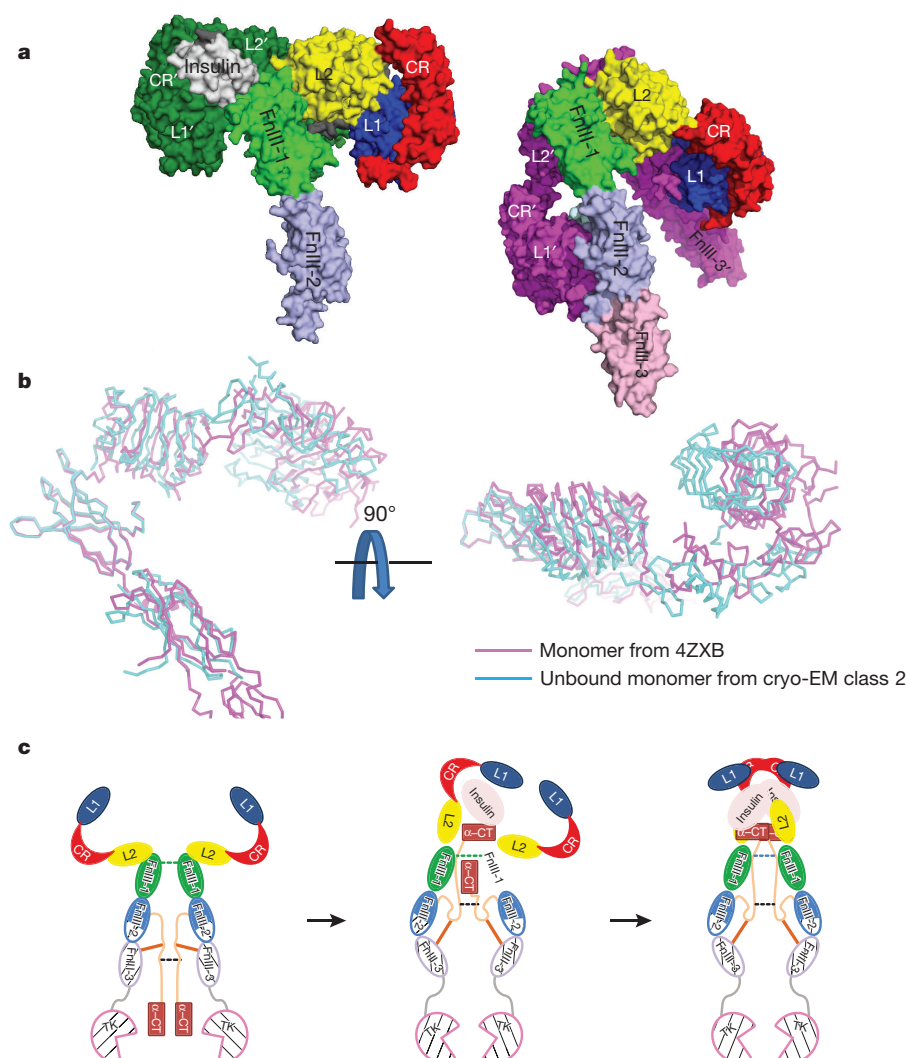


Figure 3 | Proposed transduction mechanism. **a**, Comparison between the cryo-EM (insulin bound) and the crystallographic (apo) dimers. Both dimers are shown as surface representation of the coordinates. Similar conformational differences between unbound and insulin-bound insulin receptor have been reported³⁰. **b**, Side and top views of the monomer (PDB ID: 4ZXB) overlaid onto the cryo-EM ‘open’ monomer using the FnIII-1 domain. **c**, Schematic of a possible activation mechanism for the insulin receptor. The insulin receptor subdomains, colour coded as in Fig. 1a (solid colours for the α -chain and lighter colours with thicker outer lines for the β -chain). Dotted lines, inter-monomer disulfide bonds; solid orange line, the signalling bridge²⁶. Binding of one insulin molecule to the apo receptor (left to middle panel) causes the L1–CR–L2 subdomains of one monomer, the FnIII-1 subdomain of the other and the α -CT helix

to move to generate the binding site. The movement of the α -CT helix and the attached ID- α causes, via the signalling bridge, a conformational change in the FnIII-3 domain. Because the two ID- α regions are also disulfide bonded, the movement of one is likely to be transmitted to the other, inducing a similar conformational change in the other FnIII-3 domain. These changes propagate through the transmembrane helix to the tyrosine kinase domains, inducing autophosphorylation and activation of the signalling pathway. This state is seen in the cryo-EM class 2 map. Right, binding of a second insulin molecule recruits the second α -CT (cryo-EM class 1 map) and may fully stabilize the activated complex. Although the diagram suggests that the α -CT helix involved in insulin binding is the one from the same monomer (*cis* interaction), there is no evidence, to our knowledge, that rules out a *trans* interaction.

respect to the linker connecting the L2 and FnIII-1 domains (residues Ala466–E469); the second is a 55° swing of the CR–L1 pair (as a rigid body) around the Gly306–Lys310 linker between the CR and L2 domains (not resolved in the current structure). These domain movements are similar to the rotation and translation observed upon ligand binding in the related EGFR family²⁷. On the basis of the comparison between unbound and bound receptors, it appears that the α -CT helix moves between approximately 55 Å (if we consider the helix from the same monomer) and approximately 70 Å (if we consider the α -CT from the other monomer). Figure 3c shows a schematic of a possible mechanism for insulin receptor transduction. It is likely that, following the binding of the first insulin molecule, the large shift in the position of the α -CT helix causes conformational changes in the insertion domain (ID)- α of both monomers. ID- α is disulfide bonded to FnIII-3 (Cys647–Cys860) and this connection (the ‘signalling bridge’²⁶)

may induce further conformational changes in the FnIII-3 domain, thus triggering the downstream signal propagation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 September 2017; accepted 20 February 2018.

Published online 28 February 2018.

1. Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).
2. Adamo, M., Raizada, M. K. & LeRoith, D. Insulin and insulin-like growth factor receptors in the nervous system. *Mol. Neurobiol.* **3**, 71–100 (1989).
3. Frasca, F. *et al.* The role of insulin receptors and IGF-I receptors in cancer and other diseases. *Arch. Physiol. Biochem.* **114**, 23–37 (2008).
4. Craft, S. Alzheimer disease: insulin resistance and AD—extending the translational path. *Nat. Rev. Neurol.* **8**, 360–362 (2012).

5. Seino, S., Seino, M., Nishi, S. & Bell, G. I. Structure of the human insulin receptor gene and characterization of its promoter. *Proc. Natl Acad. Sci. USA* **86**, 114–118 (1989).
6. De Meyts, P. Insulin/receptor binding: the last piece of the puzzle? What recent progress on the structure of the insulin/receptor complex tells us (or not) about negative cooperativity and activation. *BioEssays* **37**, 389–397 (2015).
7. Menting, J. G. *et al.* How insulin engages its primary binding site on the insulin receptor. *Nature* **493**, 241–245 (2013).
8. McKern, N. M. *et al.* Structure of the insulin receptor ectodomain reveals a folded-over conformation. *Nature* **443**, 218–221 (2006).
9. Smith, B. J. *et al.* Structural resolution of a tandem hormone-binding element in the insulin receptor and its implications for design of peptide agonists. *Proc. Natl Acad. Sci. USA* **107**, 6771–6776 (2010).
10. Croll, T. I. *et al.* Higher-resolution structure of the human insulin receptor ectodomain: multi-modal inclusion of the insert domain. *Structure* **24**, 469–476 (2016).
11. Whittaker, L., Hao, C., Fu, W. & Whittaker, J. High-affinity insulin binding: insulin interacts with two receptor ligand binding sites. *Biochemistry* **47**, 12900–12909 (2008).
12. Tatulian, S. A. Structural dynamics of insulin receptor and transmembrane signaling. *Biochemistry* **54**, 5523–5532 (2015).
13. De Meyts, P. & Whittaker, J. Structural biology of insulin and IGF1 receptors: implications for drug design. *Nat. Rev. Drug Discov.* **1**, 769–783 (2002).
14. Tan, Y. Z. *et al.* Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
15. Jain, T., Sheehan, P., Crum, J., Carragher, B. & Potter, C. S. Spotiton: a prototype for an integrated inkjet dispense and vitrification system for cryo-TEM. *J. Struct. Biol.* **179**, 68–75 (2012).
16. Razinkov, I. *et al.* A new method for vitrifying samples for cryoEM. *J. Struct. Biol.* **195**, 190–198 (2016).
17. Markussen, J., Halstrøm, J., Wiberg, F. C. & Schäffer, L. Immobilized insulin for high capacity affinity chromatography of insulin receptors. *J. Biol. Chem.* **266**, 18814–18818 (1991).
18. Whittaker, J. & Whittaker, L. Characterization of the functional insulin binding epitopes of the full-length insulin receptor. *J. Biol. Chem.* **280**, 20932–20936 (2005).
19. Ward, C. W., Menting, J. G. & Lawrence, M. C. The insulin receptor changes conformation in unforeseen ways on ligand binding: sharpening the picture of insulin receptor activation. *BioEssays* **35**, 945–954 (2013).
20. Hao, C., Whittaker, L. & Whittaker, J. Characterization of a second ligand binding site of the insulin receptor. *Biochem. Biophys. Res. Commun.* **347**, 334–339 (2006).
21. Roach, P. *et al.* A novel human insulin receptor gene mutation uniquely inhibits insulin binding without impairing posttranslational processing. *Diabetes* **43**, 1096–1102 (1994).
22. Accili, D., Mosthaf, L., Ullrich, A. & Taylor, S. I. A mutation in the extracellular domain of the insulin receptor impairs the ability of insulin to stimulate receptor autophosphorylation. *J. Biol. Chem.* **266**, 434–439 (1991).
23. Lebrun, C. *et al.* Antibodies to the extracellular receptor domain restore the hormone-insensitive kinase and conformation of the mutant insulin receptor valine 382. *J. Biol. Chem.* **268**, 11272–11277 (1993).
24. Kadowaki, H. *et al.* Mutagenesis of lysine 460 in the human insulin receptor. Effects upon receptor recycling and cooperative interactions among binding sites. *J. Biol. Chem.* **265**, 21285–21296 (1990).
25. Brandt, J., Andersen, A. S. & Kristensen, C. Dimeric fragment of the insulin receptor α -subunit binds insulin with full holoreceptor affinity. *J. Biol. Chem.* **276**, 12378–12384 (2001).
26. Ye, L. *et al.* Structure and dynamics of the insulin receptor: implications for receptor activation and drug discovery. *Drug Discov. Today* **22**, 1092–1102 (2017).
27. Burgess, A. W. *et al.* An open-and-shut case? Recent insights into the activation of EGF/ErbB receptors. *Mol. Cell* **12**, 541–552 (2003).
28. Schäffer, L. & Ljungqvist, L. Identification of a disulfide bridge connecting the α -subunits of the extracellular domain of the insulin receptor. *Biochem. Biophys. Res. Commun.* **189**, 650–653 (1992).
29. Sparrow, L. G. *et al.* The disulfide bonds in the C-terminal domains of the human insulin receptor ectodomain. *J. Biol. Chem.* **272**, 29460–29467 (1997).
30. Gutmann, T., Kim, K. H., Grzybek, M., Walz, T. & Coskun, U. Visualization of ligand-induced transmembrane signalling in the full-length human insulin receptor. *J. Cell Biol.* <https://doi.org/10.1083/jcb.201711047> (2018).

Supplementary Information is available in the online version of the paper.

Acknowledgements The work presented here was conducted at the National Resource for Automated Molecular Microscopy located at the New York Structural Biology Center, supported by grants from the NIH (GM103310, OD019994) and the Simons Foundation (349247). The authors would like to acknowledge the entire staff of the Simons Electron Microscopy Center at the New York Structural Biology Center for continuous help and technical support and G. Boykow, A. Ogawa and L. Zhang (*In Vitro* Pharmacology Group, Merck & Co.) for providing assay support.

Author Contributions G.S. collected and processed the data, interpreted the results and wrote the manuscript. V.P.D. prepared the grids used in the studies, collected and processed the data and assisted in writing the manuscript. Z.Z. prepared the grids used in the study. W.P. and A.H. obtained and prepared the receptor and insulin samples. T.K. produced the binding data. T.M. characterized the sample. Z.Z., W.P., A.H., T.K. and T.M. helped in writing the manuscript. C.S., C.S.P. and B.C. helped in planning the experiments, analysing the data and writing and editing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to G.S. (giovanna.scapin@merck.com).

Reviewer Information *Nature* thanks J. Rubinstein, G. Skiniotis and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Recombinant proteins. Recombinant human insulin receptor-CD220 (R&D systems, 1544-IR-50/CF) was reconstituted in PBS pH 7.4 at 0.3 mg/ml (0.00293 mM), stored in 100- μ l aliquots at -80°C and used without further purification. The protein was tested for insulin binding using ^{125}I -insulin competition and surface plasmon resonance (Extended Data Table 2). The measured binding potency and binding affinities were consistent with values from the literature^{17,31}. A Coomassie-stained SDS-PAGE gel (Extended Data Fig. 5) of the sample shows the expected bands for the α - and β -chains. The apparent molecular weights of both chains were higher than predicted (82.9 kDa for the α -subunit and 22.9 kDa for the β -subunit), and the β -subunit appears as three discrete bands. This has previously been observed^{17,32,33} and can be attributed to glycosylation (and to different states of glycosylation for the β subunit). The gel confirms the integrity of the sample used in the studies.

Recombinant human insulin (RHI) was obtained from the laboratories of Merck. RHI is a commercially available material; a generic version of the bioprocess used in preparing it has been described³⁴. A stock solution (5.6 mg/ml, 0.969 mM) was made in HBS (10 mM Hepes pH 7.4, 150 mM NaCl) and stored at 4°C .

Insulin receptor ECD competition binding assay. Purified His-tagged human insulin receptor ECD (R&D Systems) was bound to an anti-mouse IgG coated microplate via attachment to a mouse monoclonal anti-human insulin receptor capture antibody (R&D Systems). Varying concentrations of RHI were incubated with the insulin receptor ECD in the presence of 0.1 nM ^{125}I -insulin (PerkinElmer, catalogue NEX420050UC) overnight at 4°C in binding buffer (100 mM HEPES, 100 mM NaCl, 10 mM MgCl_2 , 0.02% Triton X-100, pH 8). Plate washings with binding buffer were carried out after the antibody and protein capture steps and following ligand incubation. The amount of radioligand still bound to the insulin receptor ECD was determined with either a TopCount or a MicroBeta instrument using Microscint-40. Concentration response curves were generated and IC_{50} values were calculated using a four-parameter fit.

Insulin receptor ECD direct binding assay. Purified His-tagged human insulin receptor ECD (R&D Systems) was immobilized to a Biacore CM5 chip via an anti-His antibody kit (GE Healthcare Life Sciences) following the manufacturer's instructions. RHI binding to the immobilized human insulin receptor ECD was assayed by passing varying RHI concentrations in running buffer ($1 \times$ HBS-EP; 10 mM HEPES, 150 mM NaCl, 3 mM EDTA, 0.005% P-20, pH 7.3) over the chip. General detection and data collection parameters for the Biacore T200 instrument were used to determine k_{on} , k_{off} and K_{d} values.

Grid preparation. Samples for grid preparation were generated by mixing 100 μ l insulin receptor solution and 3 μ l insulin stock (for a final insulin receptor:insulin ratio of 1:10) followed by incubating on ice for at least one hour before making vitrified grids. Poor grid reproducibility and difficulty in obtaining suitable ice thickness presented issues when using manual grid preparation. To address these issues, the samples were vitrified using a semi-automated Spotiton V1.0 robot^{15,16,35}, a device for preparing cryo-EM samples that uses piezo-electric dispensing to apply small (50-pl) drops of sample across a 'self-blotting' nanowire grid as it flies past en route to plunging into liquid ethane. The nanowire grids, manufactured in house, used either lacey carbon or lacey gold supports³⁶; the gold supports help to optimize image quality when acquiring data with the grid tilted relative to the electron beam^{14,37}. Tilted data collection was necessary to address highly preferred particle orientation observed during initial data collections. For the tilted datasets, nanowire gold substrate grids were plasma cleaned for 10 s (with O_2 and H_2). The sample was dispensed onto these grids in 50-pl drops for a total of ~ 5 nl sample dispensed in a stripe from top to bottom across each grid, before the grid was plunged into liquid ethane. The time between sample application to the grid and the plunge into ethane was typically ~ 300 ms. Subsequently, we modified the SpotItOn instrument to further reduce the time between sample application and plunge to ~ 170 ms. This modification was undertaken in an attempt to reduce the time the sample has to interact with the air-water interfaces in the thin liquid film before vitrification. These interactions are assumed to be the cause of preferred particle orientations³⁸, and indeed the faster plunging speed resulted in grids that displayed markedly better particle distribution (data not shown), enabling the collection of data at zero-degree tilt. Nanowire-based, self-blotting grids with a lacey carbon supporting substrate were used for this purpose, following the same sample preparation methods described above.

Data collection and processing. Tilted data were collected on a Titan Krios (Thermo Fisher) equipped with an energy filter and a Gatan K2 Counting camera. The microscope was operated at 300 kV and a nominal magnification of 105,000 \times , with a calibrated pixel size of 1.1 \AA . The defocus ranged from -1 to -2.5 μm . Images were collected at a tilt angle of -30° to address the issue of preferred orientation identified during initial data collection¹⁴. Exposure was set to 10 s (40 frames per image), for a total dose of $\sim 68 \text{ e}^{-} \text{\AA}^{-2}$. A total of 6,805 images was collected

in three sessions using Legion³⁹. Of these, 3,056 images were selected as suitable for further data processing. Frames were aligned using MotionCorr⁴⁰, global and per-particle CTF was calculated using gCTF⁴¹. Particle picking was done using Gautamatch (<http://www.mrc-lmb.cam.ac.uk/kzhang/>) and resulted in 1,206,222 particles. All subsequent data processing (from 2D classification to final reconstruction) was done using Cryosparc⁴². After several cycles of 3D classification, 151,409 particles in one class were used to generate a 4.6 \AA map with C2 symmetry (and a 5.6 \AA map with C1 symmetry). A C1 symmetry map was generated at 7.4 \AA resolution from 48,315 particles in a second class (Extended Data Fig. 1). Fewer than 20% of the originally selected particles contributed to the final 3D maps. This was caused firstly by very liberal initial picking criteria that resulted in a high rate of false positives that were rejected during initial 2D classification (approximately 60%). Secondly, there is a possibility that a substantial fraction of the receptor is not complexed, or alternatively, that there is a fast equilibrium between bound and unbound species that shows up as heterogeneity, limiting the contribution of these particles to the 3D structure.

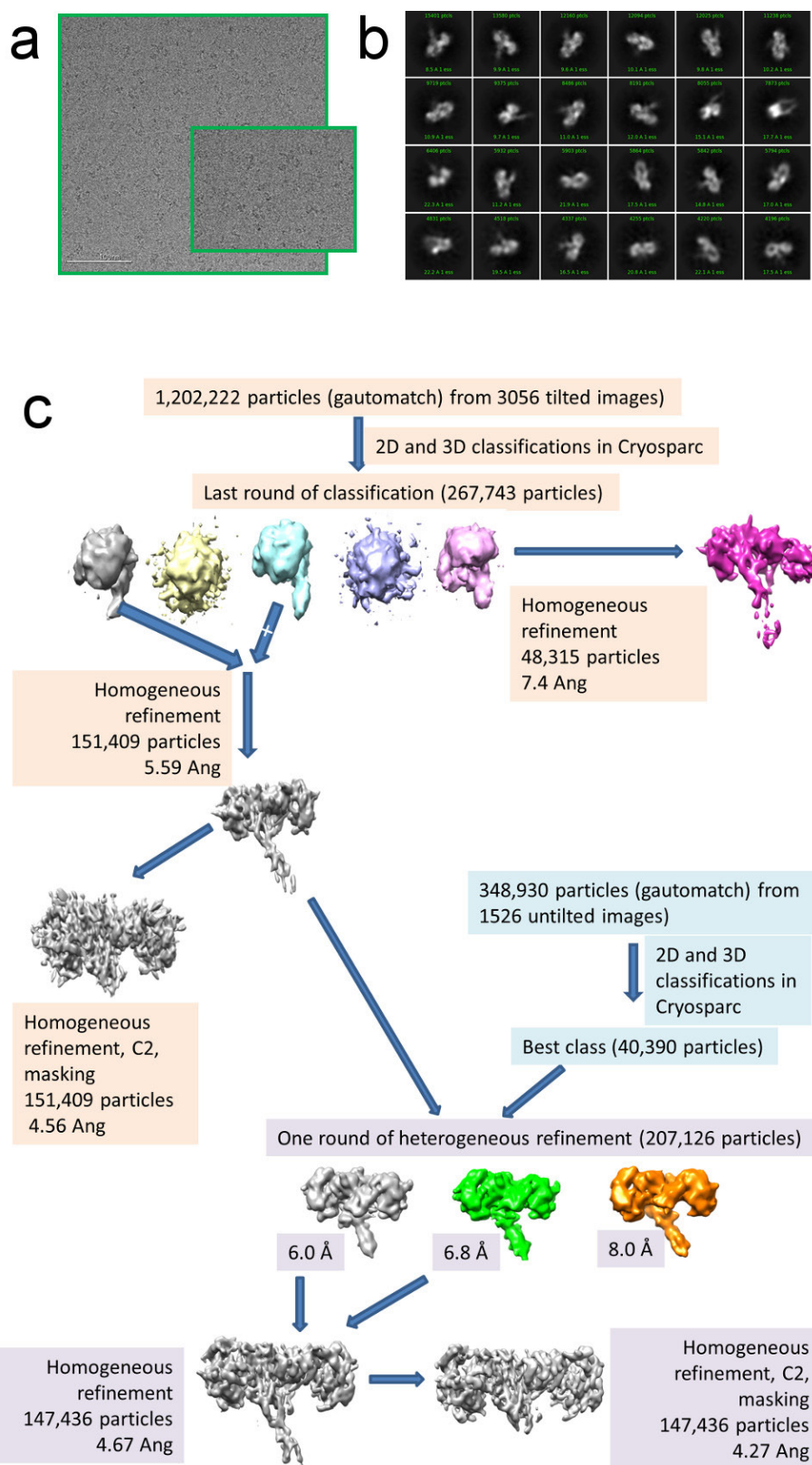
Untilted data were collected on a Titan Krios (Thermo Fisher) equipped with an energy filter and a Gatan K2 Counting camera. The microscope was operated at 300 kV and a nominal magnification of 105,000 \times , with a calibrated pixel size of 1.1 \AA . The defocus ranged from -1 to -2.5 μm . Exposure was set to 10 s (40 frames per image), for a total dose of $\sim 66 \text{ e}^{-} \text{\AA}^{-2}$. A total of 2,487 images was collected in three sessions using Legion³⁹. Of these, 1,526 images were selected as suitable for further data processing. Frames were aligned using MotionCorr⁴⁰, global and per-particle CTF was calculated using gCTF⁴¹. Particle picking was done using Gautamatch (<http://www.mrc-lmb.cam.ac.uk/kzhang/>) and produced 348,930 particles. All subsequent data processing (from 2D classification to final reconstruction) was done using Cryosparc⁴². After several cycles of 3D classification, a set of 40,390 particles was identified, which produced a 4.7 \AA class 1 map after symmetrized homogeneous refinement. This set was merged with the 151,409 particles from the tilted class 1 set. After a round of heterogeneous refinement, 147,436 particles were used to generate a 4.3 \AA map with C2 symmetry (and a 4.7 \AA map with C1 symmetry) (Extended Data Fig. 1). The resolution of each reconstruction was determined by the gold standard Fourier shell correlation (FSC) criterion in Cryosparc⁴² and RELION PostProcess⁴³. The completeness of the maps was assessed by measuring the Euler angle orientation distribution (as calculated by Cryosparc⁴²) and by evaluating the 3D FSC¹⁴ of each map. (Extended Data Figs 2, 3, 6). The quality of the map was confirmed by the ability to see separated β -strands and the presence of bulky side chains (Extended Data Fig. 2f).

Structure solution. Individual domains of the insulin receptor (L1, CR, L2 and FnIII-1, from PDB ID: 4XZB) were positioned into the initial 4.6 \AA resolution map using Molrep⁴⁴ and rigid-body refined using Coot⁴⁵. The insulin and the α -CT helix were positioned by overlaying the microreceptor structure from PDB ID 3W11⁹ onto the L1 domain of the complex and rigid body fitting them into the available density. The α -CT helix was manually extended, while the full insulin molecule was built using as a reference one of the monomers from PDB ID 1ZNI. The full structure was subjected to five cycles of global real-space refinement with NCS, rotamer, Ramachandran plot and C-beta deviations restraints enabled in Phenix⁴⁶. This model was subsequently refined against the 4.3 \AA map following the same real-space refinement procedure in Phenix⁴⁶. The resulting model is identical to the 4.6 \AA model, but for the addition of the first seven N-terminal residues (His1-Val7) to both receptor monomers, and contains residues His1-Asp591 and Lys691-Val720 (α -CT helix) of both of insulin receptor monomers and residues Gly1-Lys21 of chain A and Phe1-Ala30 of chain B of both bound insulins. Sugars were added to residues Asn16, Asn25, Asn111, Asn255, Asn397 and Asn418 of both insulin receptor monomers. The sugars were extracted from PDB ID 4ZXB. All subsequent, lower resolution models were generated by manually positioning the higher resolution structure onto the available density and rigid-body refining the individual domains to their final position using Coot⁴⁵. Subsequently, the structures were subjected to five cycles of global real-space refinement with rotamer, Ramachandran plot and C-beta deviations restraints enabled in Phenix⁴⁶. The lack of definition for the FnIII-2 and especially FnIII-3 domains may be due to the intrinsic flexibility of these regions. While one FnIII-2 region was clearly visible in the un-symmetrized map, unstructured density that could be attributed to the remaining missing regions can be seen by contouring the maps at very low sigma (Extended Data Fig. 5).

Figure 2a and Extended Data Figs 1c, 2e, 3e, 6e were generated with Chimera⁴⁷. All other figures, unless otherwise specified, were generated with PyMOL (The PyMOL Molecular Graphics System v1.8).

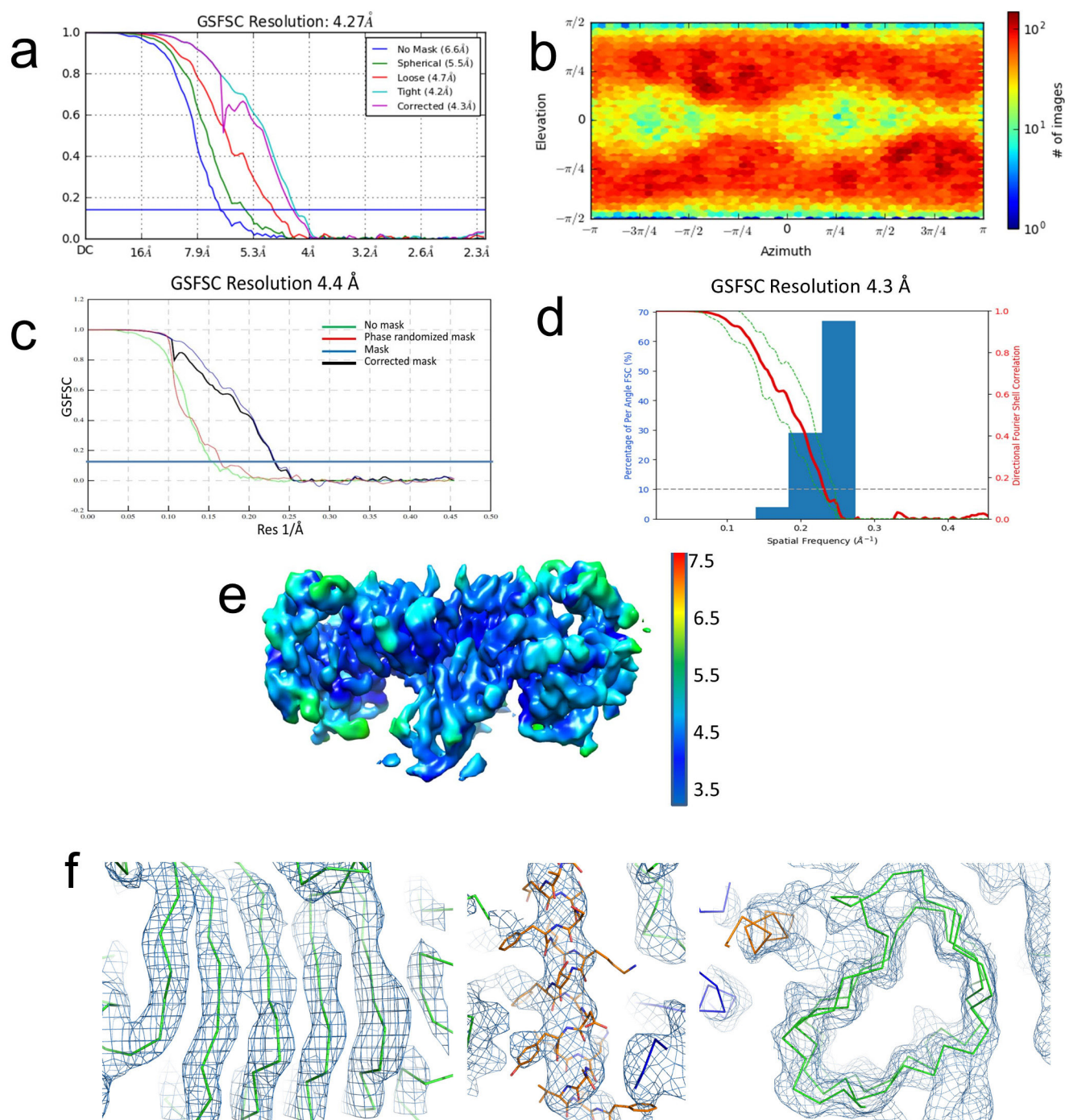
Data availability. The maps and coordinates generated and analysed during the current study have been deposited with the Electron Microscopy Data Bank and Protein Data Bank under accession codes EMD-7462 (PDB 6CE9), EMD-7463 (PDB 6CEB) and EMD-7461 (PDB 6CE7) for the 4.3 \AA , 4.7 \AA and 7.4 \AA maps, respectively.

31. Subramanian, K., Fee, C. J., Fredericks, R., Stubbs, R. S. & Hayes, M. T. Insulin receptor–insulin interaction kinetics using multiplex surface plasmon resonance. *J. Mol. Recognit.* **26**, 643–652 (2013).
32. Johnson, J. D., Wong, M. L. & Rutter, W. J. Properties of the insulin receptor ectodomain. *Proc. Natl Acad. Sci. USA* **85**, 7516–7520 (1988).
33. Cosgrove, L. *et al.* Purification and properties of insulin receptor ectodomain from large-scale mammalian cell culture. *Protein Expr. Purif.* **6**, 789–798 (1995).
34. Johnson, I. S. Human insulin from recombinant DNA technology. *Science* **219**, 632–637 (1983).
35. Dandey, V. P. *et al.* Spotiton: new features and applications. *J. Struct. Biol.* <https://dx.doi.org/10.1016/j.jsb.2018.01.002> (2018).
36. Wei, H. *et al.* Optimizing “self-wicking” nanowire grids. *J. Struct. Biol.* <https://dx.doi.org/10.1016/j.jsb.2018.01.001> (2018).
37. Russo, C. J. & Passmore, L. A. Electron microscopy: ultrastable gold substrates for electron cryomicroscopy. *Science* **346**, 1377–1380 (2014).
38. Glaeser, R. & Han, B. Opinion: hazards faced by macromolecules when confined to thin aqueous films. *Biophys. Rep.* **3**, 1–7 (2017).
39. Suloway, C. *et al.* Automated molecular microscopy: the new Legimon system. *J. Struct. Biol.* **151**, 41–60 (2005).
40. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
41. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
42. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
43. Scheres, S. H. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.* **415**, 406–418 (2012).
44. Vagin, A. & Teplyakov, A. I. MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* **30**, 1022–1025 (1997).
45. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
46. Afonine, P., Headd, J., Terwilliger, T. & Adams, P. New tool: phenix.real_space_refine. *Comput. Crystallogr. Newsl.* **4**, 43–44 (2013).
47. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
48. Menting, J. G. *et al.* Protective hinge in insulin opens to enable its receptor engagement. *Proc. Natl Acad. Sci. USA* **111**, E3395–E3404 (2014).

**Extended Data Figure 1 | Data collection and processing.**

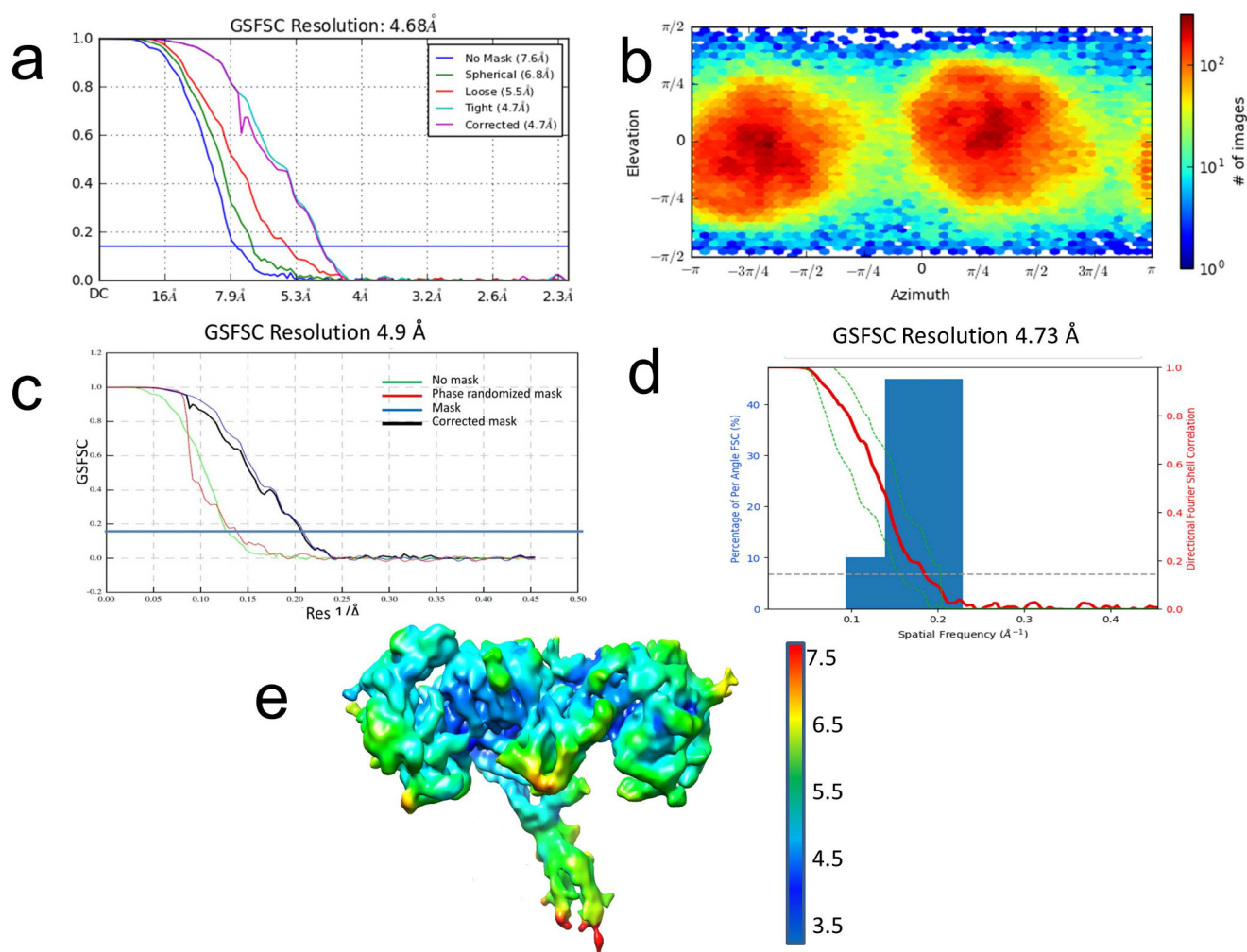
a, Representative micrograph of the insulin receptor–insulin complex. Images were collected on a Krios (Thermo Fisher) equipped with an energy filter and a Gatan K2 Counting camera; the magnification was

set to 105,000 \times , with a calibrated pixel size of 1.10 Å. **b**, Representative 2D class averages as calculated with Cryosparc⁴². **c**, Schematic of data processing.



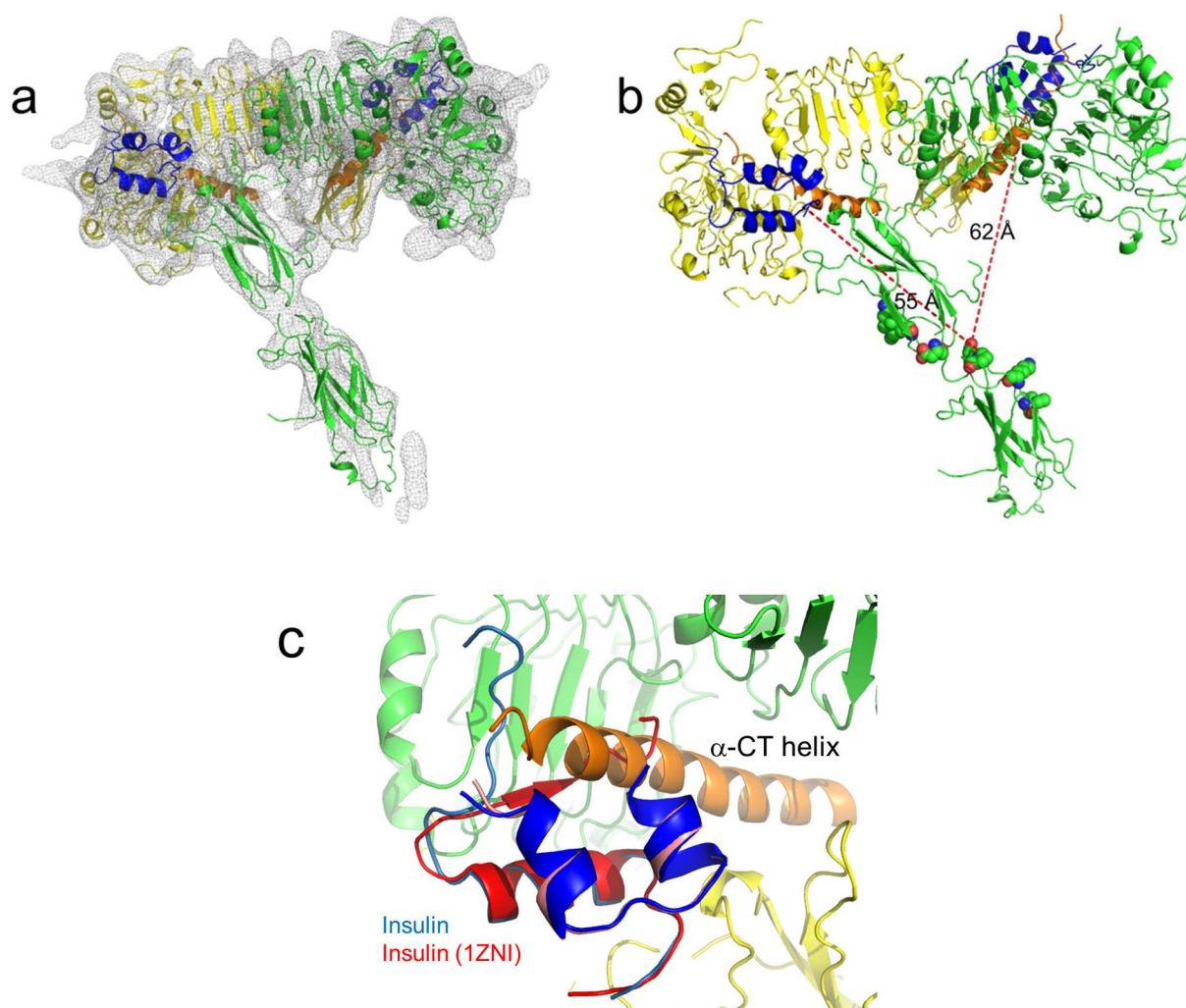
Extended Data Figure 2 | Map of class 1 reconstructed using C2 symmetry to a resolution of 4.3 Å. **a**, Gold standard FSC curve from Cryosparc⁴². **b**, Euler angle orientation distribution from Cryosparc⁴². **c**, Gold standard FSC curve as calculated in RELION⁴³. **d**, Plot of the global half-map FSC (solid red line) and spread of directional resolution values ($\pm 1\sigma$ from mean, green dotted lines; the blue bars are a histogram

of 100 such values evenly sampled over the 3D FSC¹⁴). **e**, Local resolution distribution (as calculated in Cryosparc⁴²). **f**, Selected areas from the class 1 C2 map at a resolution of 4.3 Å (map contoured at 12σ). Left, individual β-strands in the L2 region are well separated; middle, bulky side chains are visible in the electron density; right, density for the L1 β-barrel.



Extended Data Figure 3 | Map of class 1 reconstructed using C1 symmetry to a resolution of 4.7 Å. a, Gold standard FSC curve from Cryosparc⁴². **b,** Euler angle orientation distribution from Cryosparc⁴². **c,** Gold standard FSC curve as calculated in RELION⁴³. **d,** Plot of the

global half-map FSC (solid red line) and spread of directional resolution values ($\pm 1\sigma$ from mean, green dotted lines; the blue bars are a histogram of 100 such values evenly sampled over the 3D FSC¹⁴). **e,** Local resolution distribution (as calculated in Cryosparc⁴²).

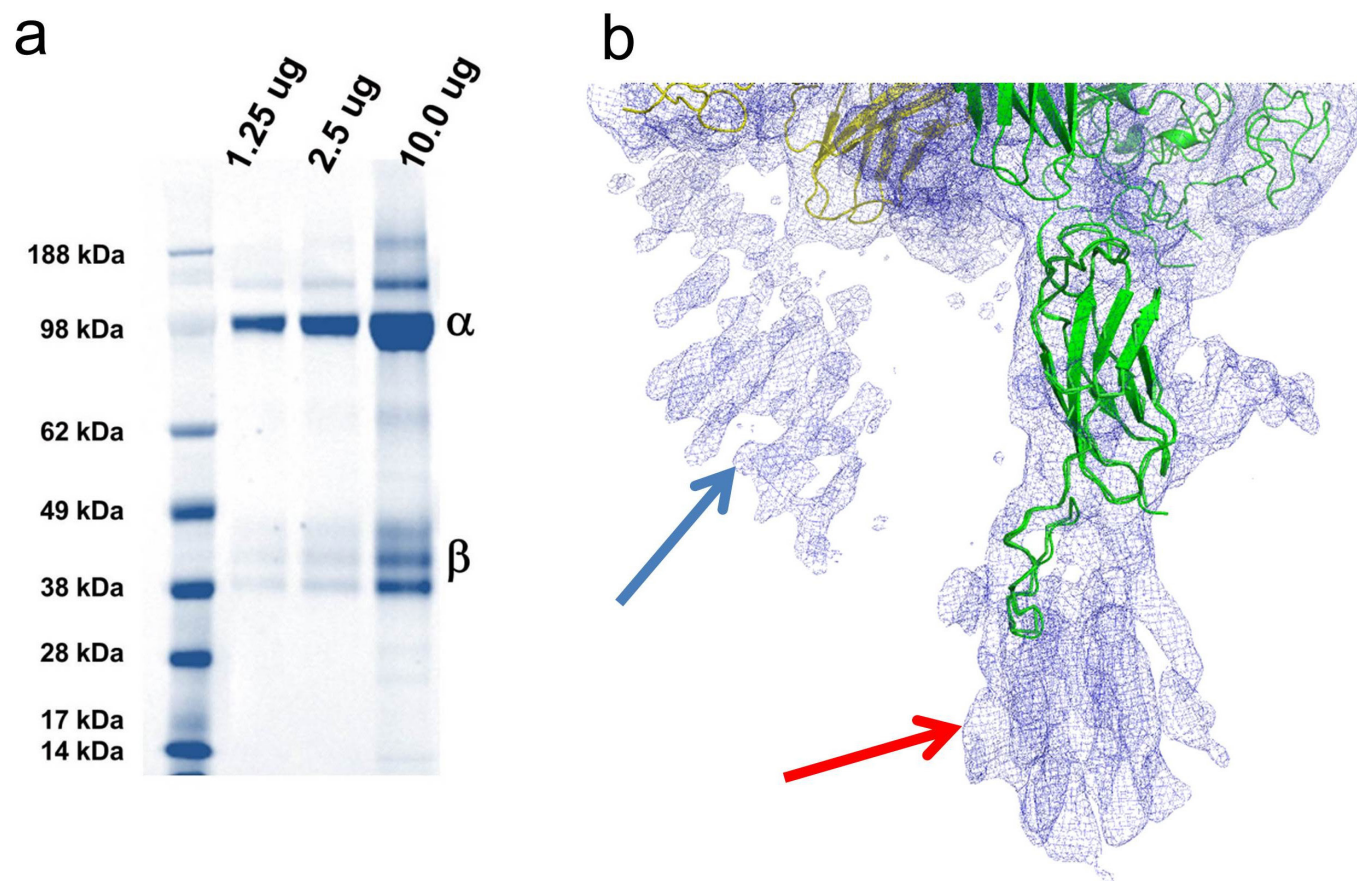


Extended Data Figure 4 | Insulin binding and crystallographic S2 site.

a, Map of the class 1 structure obtained without applying C2 symmetry, with the insulin receptor model shown as a cartoon. The map is asymmetric and only one of the FnIII-2 subdomains is clearly visible.

b, Positioning of the FnIII-2 subdomain enables analysis of the relative

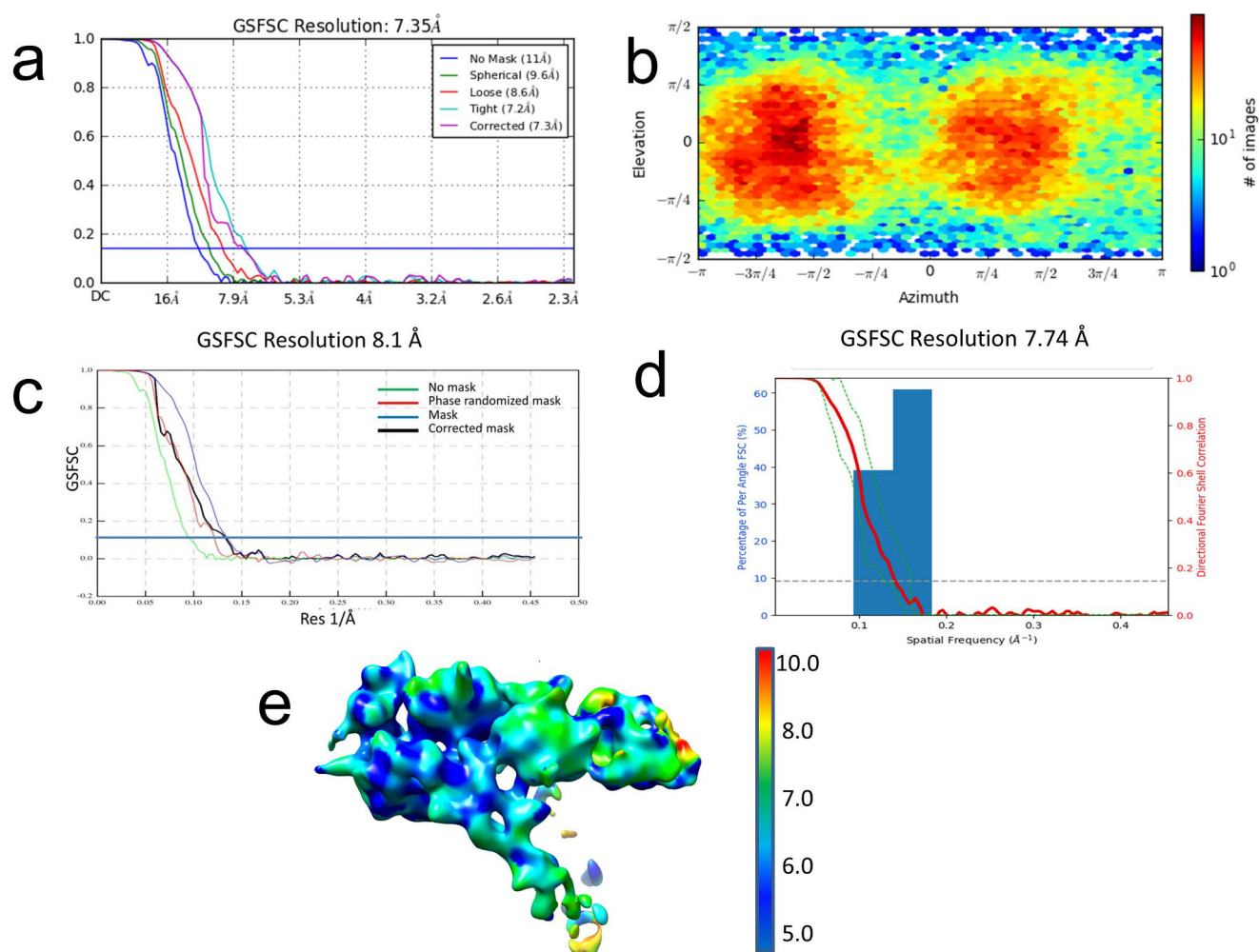
position of the bound insulin (blue) and the proposed S2 site (shown as spheres, residues selected according to ref. 26: the insulin is between 55 and 62 Å away, and on the opposite side of the proposed S2). **c**, Overlay of free insulin (PDB ID: 1ZNI) onto the insulin bound to the insulin receptor.



Extended Data Figure 5 | Insulin receptor sample characterization.

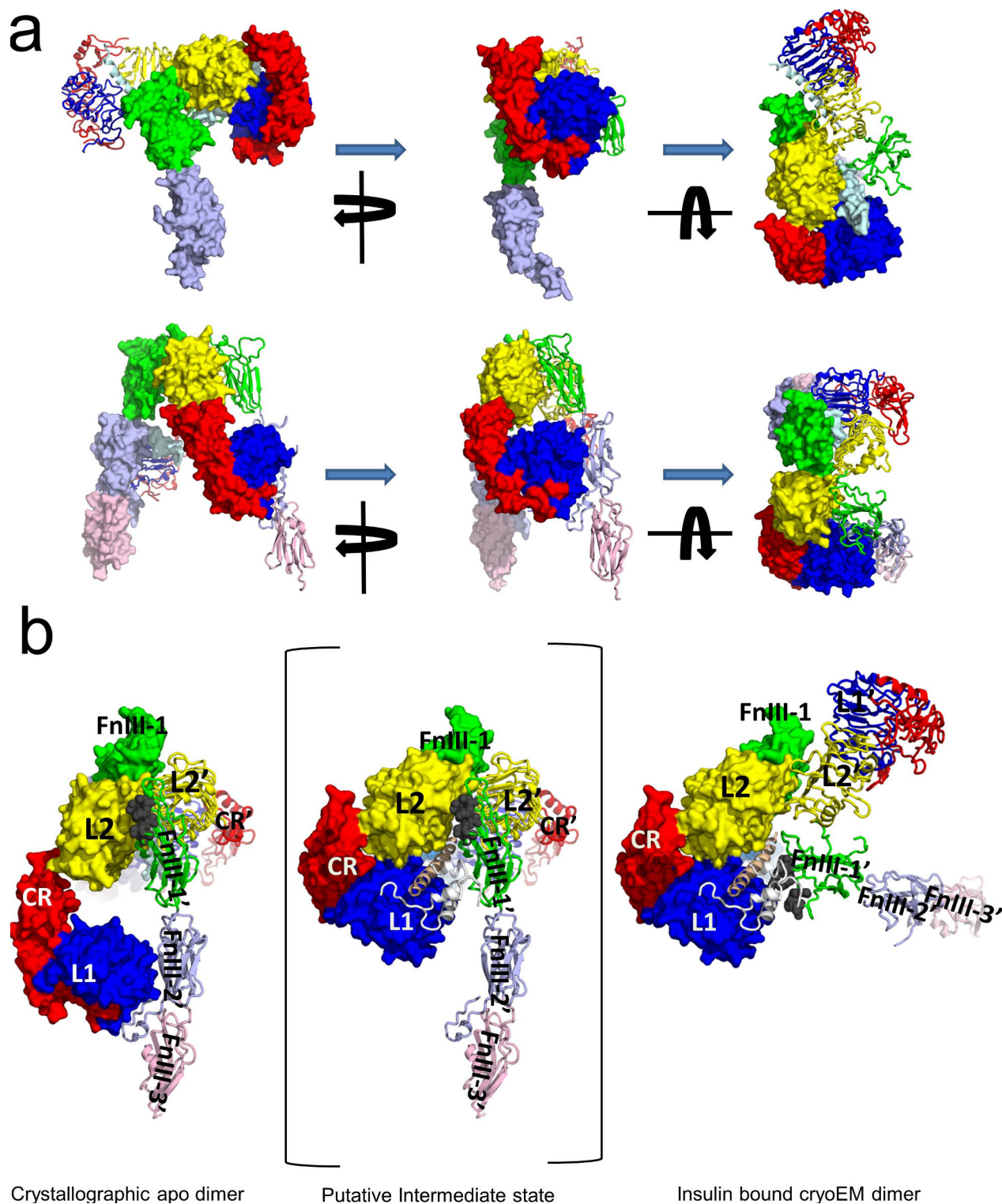
a, Coomassie-stained SDS-PAGE gel. The protein was solubilized in PBS at pH 7.2 and run on a 4–12% Bis-Tris gel in 1× MOPS. Molecular weight markers are labelled. Bands corresponding to the α - and β -chains are indicated. This experiment was only run once to confirm sample quality

was as reported by R&D Systems. For gel source data, see Supplementary Fig. 1. **b**, Cryo-EM density for the un-symmetrized map (4.7 Å resolution) countered at 6σ . Density that can be attributed to the second FnIII-2 domain (blue arrow) as well as one of the FnIII-3 domains (red arrow) is visible, but it is not of sufficient quality for building the model.



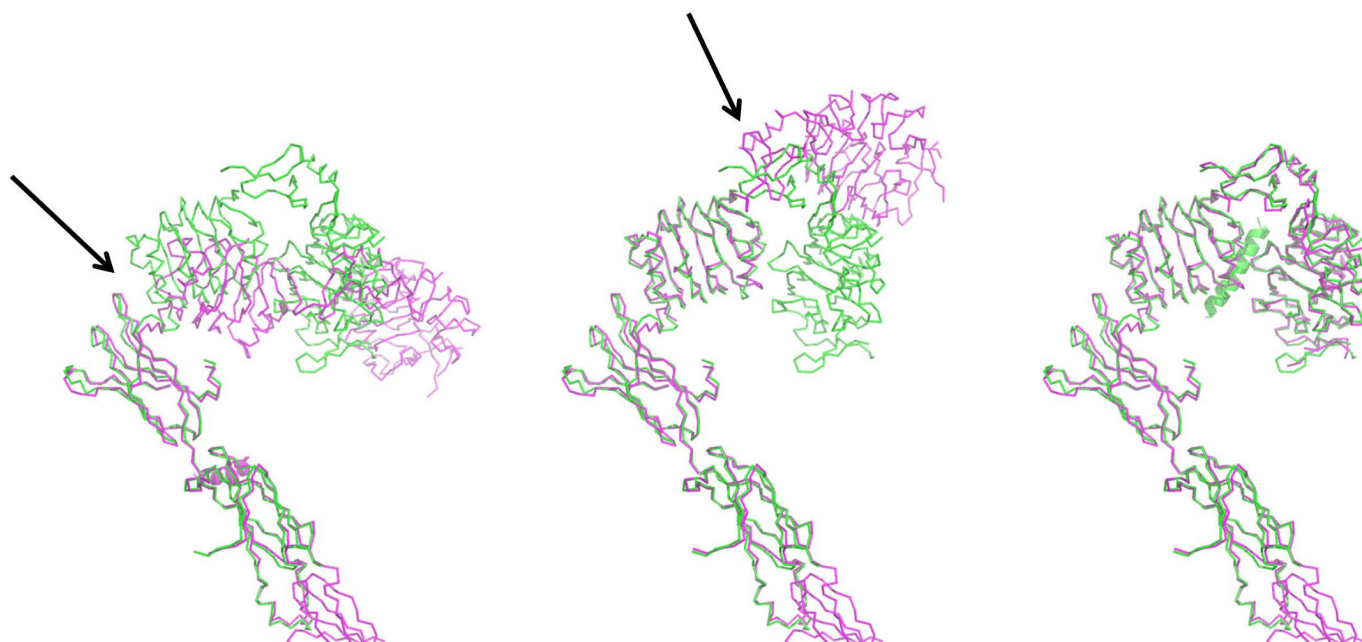
Extended Data Figure 6 | Map of class 2 reconstructed using C1 symmetry to a resolution of 7.4 Å. **a**, Gold standard FSC curve from Cryosparc⁴². **b**, Euler angle orientation distribution from Cryosparc⁴². **c**, Gold standard FSC curve as calculated in RELION⁴³. **d**, Plot of the

global half-map FSC (solid red line) and spread of directional resolution values ($\pm 1\sigma$ from mean, green dotted lines; the blue bars are a histogram of 100 such values evenly sampled over the 3D FSC¹⁴). **e**, Local resolution distribution (as calculated in Cryosparc⁴²).



Extended Data Figure 7 | Comparison between the crystallographic and the cryo-EM dimer. **a**, The top panel shows three different views of the cryo-EM dimer, related by the 90° rotation shown in the figure. These are surface (for one monomer) and cartoon (for the other) representations of the coordinates fitted to the un-symmetrized, 4.7 Å map. In this model one of the FnIII-2 subdomains is visible. Both monomers are coloured according to the subdomains (L1, blue; CR, red; L2, yellow; FnIII-1, green; FnIII-2, light blue). The bottom panel shows three corresponding views for the crystallographic dimer. One monomer is shown as surface, the other as a cartoon; both are coloured according to the subdomains (in the cryo-EM dimer, the additional FnIII-3 is coloured in light purple). The relative

arrangements of the two monomers in the cryo-EM and crystallographic dimers are completely different. **b**, If the crystallographic insulin receptor dimer (left panel) represents the biological apo insulin receptor dimer, the transition to the insulin-bound dimer (right panel) would imply a putative intermediate state, shown in the middle panel. In this state, the large conformational changes required to accommodate the α -CT helix (orange ribbon) and insulin (white ribbon), and to allow them to engage the S2 site (black spheres) would require disruption of the extensive surface interface between L1 and FnIII-2' and 3' and L2-FnIII-1' interactions. In the middle and right panels the α -CT helix and the insulin are from the cryo-EM structure.



Extended Data Figure 8 | Comparison between the bound and unbound insulin receptor monomer. Transition from the apo (magenta, PDB ID: 4ZXB monomer) to the bound form (green, cryo-EM structure) for

the insulin receptor monomer. This transition can be described as two rotations with respect to the linker regions identified by the black arrows. The α -CT helix is shown as a cartoon.

Extended Data Table 1 | Summary of the interactions between the insulin receptor and bound insulin

Insulin residues	IR L1 (IR monomer A)	IR α -CT helix	IR L2 (IR monomer A)	IR FnIII-1' (IR monomer B)
	L62, F64, F88, F89, Y91, V94, F96, R188	F701, F705		
	R14, Q34, L36, F88	Y708		
	R118	E698		
	E120	R702		
	R37, F64	L709		
V3A, G8B, L11B, V12B		D707, H710		
G1A, I2A, Y19A, L15B		F714		
V12B	F39, F64, R65			
Y16B	F39, K40			
F24B	R14, N15, L37	V713		
F25B		P716, S717, S719		
Y26B	D12, R14	P716		
Chain A, C7-T8; Chain B, E4-H10*				P495-R498, R539- N541, E575
		L696-K703	G346-N349, R372- Y374	
		L696-K703		D496, R498-L501, V570-T571

Interactions are shown along the rows. Light blue: structurally characterized interactions⁷; red: interactions predicted by biochemical analysis⁶, not seen in the initial microreceptor structures⁷, but subsequently visualized in another study⁴⁸; green: new interactions identified in the cryo-EM map.

*Insulin residues predicted to belong to S2, with no IR partner assigned⁶.

Extended Data Table 2 | Human insulin receptor ECD binding and kinetic data

[¹²⁵I]-Insulin Competition Binding: IC₅₀ (nM)	Biacore Binding: High Affinity Site K_d (nM)	Biacore Binding: Low Affinity Site K_d (nM)
3.0	2.7	220
(2.6)	(1.2)	(1.7)

Biacore Binding: High Affinity Site k_{on} (nM⁻¹ s⁻¹)	Biacore Binding: High Affinity Site k_{off} (s⁻¹)	Biacore Binding: Low Affinity Site k_{on} (nM⁻¹ s⁻¹)	Biacore Binding: Low Affinity Site k_{off} (s⁻¹)
0.0016	0.0044	0.00061	0.13
(1.1)	(1.1)	(1.3)	(1.3)

Top, human insulin receptor ECD binding data for recombinant human insulin. Data are reported as the geometric mean for 23 replicates; the geometric standard deviation is shown in parentheses. The binding potency measured with the ¹²⁵I-insulin competition assay agrees very well with values from the literature¹⁷. Bottom, human insulin receptor ECD binding kinetic data for RHI. Data are reported as the geometric mean of three replicates; the geometric standard deviation is shown in parentheses. The surface plasmon resonance data are consistent with values from the literature³¹.

Modular assembly of the nucleolar pre-60S ribosomal subunit

Zahra Assur Sanghai^{1*}, Linamarie Miller^{1,2*}, Kelly R. Molloy³, Jonas Barandun¹, Mirjam Hunziker¹, Malik Chaker-Margot^{1,2}, Junjie Wang³, Brian T. Chait³ & Sebastian Klinge¹

Early co-transcriptional events during eukaryotic ribosome assembly result in the formation of precursors of the small (40S) and large (60S) ribosomal subunits¹. A multitude of transient assembly factors regulate and chaperone the systematic folding of pre-ribosomal RNA subdomains. However, owing to a lack of structural information, the role of these factors during early nucleolar 60S assembly is not fully understood. Here we report cryo-electron microscopy (cryo-EM) reconstructions of the nucleolar pre-60S ribosomal subunit in different conformational states at resolutions of up to 3.4 Å. These reconstructions reveal how steric hindrance and molecular mimicry are used to prevent both premature folding states and binding of later factors. This is accomplished by the concerted activity of 21 ribosome assembly factors that stabilize and remodel pre-ribosomal RNA and ribosomal proteins. Among these factors, three Brix-domain proteins and their binding partners form a ring-like structure at ribosomal RNA (rRNA) domain boundaries to support the architecture of the maturing particle. The existence of mutually exclusive conformations of these pre-60S particles suggests that the formation of the polypeptide exit tunnel is achieved through different folding pathways during subsequent stages of ribosome assembly. These structures rationalize previous genetic and biochemical data and highlight the mechanisms that drive eukaryotic ribosome assembly in a unidirectional manner.

The assembly of the large eukaryotic ribosomal subunit (60S) is organized as a series of consecutive intermediates, which are regulated by different sets of ribosome assembly factors in the nucleolus, nucleus and cytoplasm¹. Three major types of pre-60S particles have been observed, each with a distinct composition of assembly factors, representing late nucleolar², nuclear³ and late nucleocytoplasmic intermediates⁴. The earliest pre-60S particles exist in the nucleolus, where they undergo major RNA processing steps and conformational changes. Subsequently, in the nucleoplasm, the internal transcribed spacer 2 (ITS2) RNA is removed and the particles are assembled into a nuclear-export-competent state. Exported particles then complete the final steps of maturation in the cytoplasm.

While previous high-resolution structural studies have revealed the architectures of the late nucleolar and export-competent pre-60S particles, the architecture of the early nucleolar particles, containing specific factors such as Nsa1, remains unknown^{2,4}. Although the identities and approximate binding regions of many early nucleolar ribosome assembly factors are known, their structures and functions have not yet been determined⁵.

To elucidate the mechanisms that govern early nucleolar large-subunit assembly, several laboratories, including ours (data not shown), have observed that cellular starvation extends the lifetime of a 27SB pre-rRNA-containing species^{6,7}. Here, we have isolated these intermediates from starved yeast using tandem-affinity purification involving tagged ribosome assembly factors Nsa1 and Nop2 (Extended Data Fig. 1).

Similar to the small subunit processome, the nucleolar pre-60S particle, which is compositionally related to Nsa1-containing particles⁸, accumulates upon starvation. We analysed purified nucleolar pre-60S particles by cryo-EM, revealing a high-resolution core (3.4 Å) that was further sub-classified into three states, which we structurally resolved at resolutions of 4.3 Å (state 1), 3.7 Å (state 2) and 4.6 Å (state 3) (Table 1, Extended Data Figs 2, 3). These three states, together with cross-linking and mass spectrometry data, resulted in the identification of 21 ribosome assembly factors (Extended Data Table 1 and Supplementary Dataset 1). Atomic models could be completed for 18 of these proteins, and homology and poly-alanine models were used for proteins such as Ebp2, Mak11 and Ytm1, which were located in more flexible regions (Fig. 1, Extended Data Fig. 4).

The purified pre-60S particles contain 27SB rRNA that has not yet been processed at ITS2 (Extended Data Fig. 1). We observed three different conformational states of this RNA (Extended Data Figs 2, 5). State 1 includes ordered density for ITS2, domains I and II, and the 5.8S rRNA. State 2 additionally revealed density for domain VI, which is present in a near-mature conformation. In contrast to state 2, state 3 lacks an ordered domain VI but features domain III. Although present, the majority of domains IV and V and the 5S ribonucleoprotein (RNP) were poorly resolved in all of the reconstructions, owing to conformational flexibility. Low-resolution features corresponding to parts of domain V (helices 74–79) and its proximal assembly factor Mak11 can be seen in states 2 and 2A (Fig. 1, Extended Data Fig. 2).

A striking feature of the nucleolar pre-60S particle is its open architecture, in which the solvent-exposed domains I, II and VI are encapsulated by a series of ribosome assembly factors as visualized in state 2 (Fig. 1, Extended Data Fig. 5, Supplementary Video). Notably, ribosomal proteins that have been associated with Diamond-Blackfan anaemia⁹ are located at critical rRNA domain interfaces in the structure (Extended Data Fig. 6), suggesting that their architectural roles are especially important during early nucleolar assembly, in which defects can trigger the nucleolar stress response.

Domains I, II and VI adopt an open conformation that is chaperoned by eight early ribosome assembly factors, which form a ring-like structure at the solvent-exposed side (Fig. 2a). In particular, Brix-domain-containing factors (Brx1, Rpf1 and Ssf1) act in conjunction with their respective binding partners (Ebp2, Mak16 and Rrp15) to interconnect these junctions and sterically prevent premature RNA–protein and RNA–RNA contacts. Architectural support for the major interface between domains I and II is provided by Rpf1 and its zinc-binding interaction partner Mak16, the helical repeat protein Rrp1 and the beta-propeller Nsa1. Rpf1 and Nsa1 occupy a region near the domain I binding site of Rpl17, while Mak16 and Rrp1 interface predominantly with ribosomal proteins Rpl4 and Rpl32 within domain II (Fig. 2b).

The ring-like structure encapsulating domains I, II and VI is continued in one direction by the Ssf1–Rrp15 heterodimer and Rrp14.

¹Laboratory of Protein and Nucleic Acid Chemistry, The Rockefeller University, New York, New York 10065, USA. ²Tri-Institutional Training Program in Chemical Biology, The Rockefeller University, New York, New York 10065, USA. ³Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, The Rockefeller University, New York, New York 10065, USA.

*These authors contributed equally to this work.

Table 1 | Cryo-EM data collection parameters and refinement and validation statistics

	Core	State 2 EMD-7324 PDB 6C0F	State 2A	State 3 EMD-7445 PDB 6CB1
Data collection and processing				
Magnification	22,500×			
Voltage (kV)	300			
Pixel size (Å)	1.3			
Electron exposure (e ⁻ Å ⁻²)	47			
Defocus range (μm)	1.0–3.5			
Symmetry imposed	C1			
Initial particle images	1,653,290			
Final particle images	514,746	201,114	75,512	31,419
Resolution (Å)	3.4	3.7	4.2	4.6
FSC threshold	0.143			
Map sharpening B-factor (Å ²)	−68.7	−71.7	−83.0	−94.2
Refinement				
Model composition				
Non hydrogen atoms		91,741		69,945
Protein residues		7,197		6,347
RNA bases		1,602		1,615
Ligands		3		2
r.m.s.d.				
Bond length (Å)		0.007		0.006
Angles (°)		1.11		1.10
Validation				
MolProbity score		1.91		1.68
Clashscore		10.67		5.96
Rotamer outliers (%)		0.30		0.13
Good sugar puckers (%)		98		97
Ramachandran				
Favoured (%)		94.75		94.85
Allowed (%)		5.18		5.02
Outliers (%)		0.07		0.13

While the long C-terminal helix of Rrp14 bridges domains II and VI, the Ssf1–Rrp15 complex is positioned at the interface of domains I and VI (Fig. 2a). Here, Ssf1 occupies the same position as Rpl31, which in the later Nog2-containing pre-60S particles binds at the interface of domains III and VI near the polypeptide exit tunnel (PET)². The PET, which is created by domains I, III and VI at the solvent-exposed side, is already formed in the Nog2 particle, where

it is blocked by the C-terminal domain (CTD) of the GTPase Nog1 (Fig. 2c, d).

The role of the Brx1–Ebp2 heterodimer near the interface of domains I and II is twofold. As well as being involved in the stabilization of these domains, its strategic binding site prevents the premature assembly of the large subunit by steric hindrance. In later stages of large subunit assembly, an RNA segment of domain I (helix 22) base-pairs with a region in domain V (helix 88) near a separate region of domain IV (helix 68) (Fig. 2e, f). Brx1 remodels helix 22 of domain I to block the premature formation of this tertiary structure with helix 88. Similarly, Brx1 prevents the mature conformations of domain IV (helix 68), domain II (expansion segment 9) and the C-terminal region of Rpl13 in this region (Fig. 2e, f).

In state 3, we identified the Erb1–Ytm1 heterodimer bound to domain III via Rpl27 (Fig. 3a). The N-terminal region of Erb1 (residues 239–397) wraps around the entire ITS2–domain I interface and is positioned underneath Nop16 and Has1 (Fig. 3b, c). This location is consistent with previous cross-linking data and explains why deletions in this region prevent the incorporation of Erb1 into pre-60S particles^{10,11}. Nop16 interconnects RNA elements of the 5.8S rRNA and regions of domain I. Additionally, Nop16 interacts with both Rpl8 and Rpl13 (Fig. 3c). The DEAD-box helicase Has1 is positioned at the interface of Rpl8, Cic1, Nop16 and Erb1 (Fig. 3c). The assembly factors Cic1, Rlp7, Nop7 and Nop15 appear in both the nucleolar pre-60S particle and the Nog2 particle in largely the same conformation (Fig. 3c, d).

Notably, the N-terminal segment of Erb1 employs molecular mimicry by binding to Nop7 in a similar fashion as Nop53 in the Nog2 particle, which uses a structurally related motif to bind to Nop7. The resulting steric hindrance is exacerbated by the alternate conformation of the N terminus of Rlp7, which further prevents Nop53 binding (Fig. 3e, f). Therefore, the coordinated mechanical removal of Erb1 and its proximal factors Ytm1, Nop16 and Has1 by Mdn1 is required before Nop53 can bind to the Nog2 particle and recruit the exosome-associated RNA helicase Mtr4 for ITS2 processing^{12,13}. The Has1 helicase may have acted upon its substrate at an earlier stage during the 27SA₃-to-27SB transition. Alternatively, it may remodel flexible RNA elements in its vicinity for the ensuing 27SB processing¹⁴.

The nucleolar pre-60S states 2 and 3 represent distinct assembly intermediates of the polypeptide exit tunnel (Fig. 4). Ssf1, Rrp15 and Rrp14 are ordered in state 2, where they chaperone domains I and

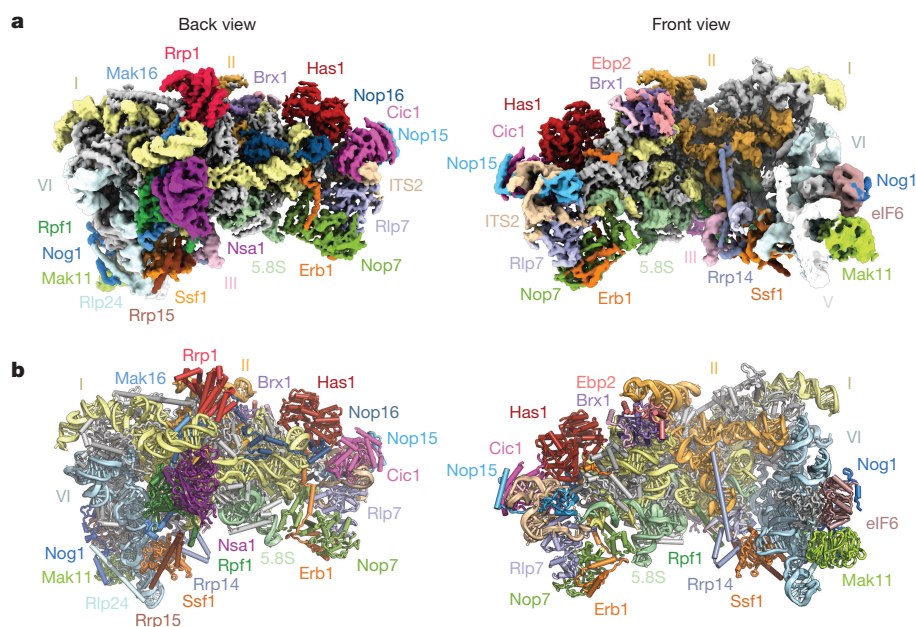


Figure 1 | Structure of the early nucleolar pre-60S particle. a, Composite cryo-EM density map of state 2, consisting of 25S rRNA domains I and II (3.4 Å) and VI (3.7 Å) (low-pass filtered to 5 Å), and associated proteins.

b, Corresponding near-atomic model of state 2 with ribosome assembly factors and 25S rRNA domains labelled and colour-coded. Ribosomal proteins are shown in grey.

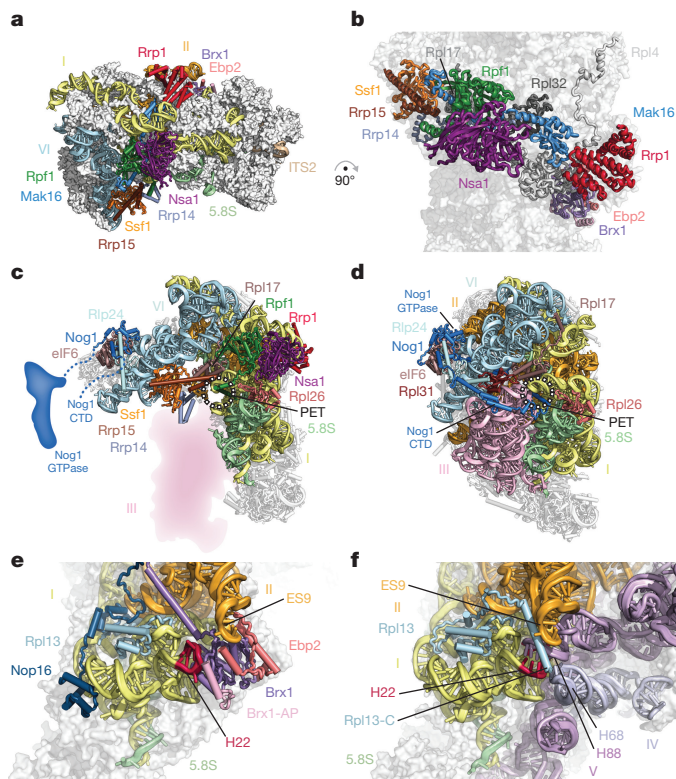


Figure 2 | A ring of nucleolar assembly factors prevents premature folding of the 25S rRNA. **a, b**, Assembly factors chaperone areas of domains I, II and VI and interact with ribosomal proteins. **c, d**, Assembly factors prevent the formation of the polypeptide exit tunnel (PET). **c**, In state 2, Ssf1–Rrp15 blocks the binding of Rpl31. Nog1 is largely unstructured. **d**, In the Nog2 particle (PDB 3JCT), the PET is formed, probed by Nog1 and supported by Rpl31. **e, f**, Brx1–Ebp2 and an associated peptide (Brx1–AP) remodel domain I (helix 22) to prevent binding of domain V (helix 88) and domain IV (helix 68) (**e**); these interactions are present in the Nog2 particle (**f**).

VI, which line two sides of the forming PET (Fig. 2b, c). By contrast, domain VI, Ssf1, Rrp15 and Rrp14 are disordered in state 3. Here, Ytm1 and Erb1 chaperone domain III, which adopts a mature conformation with respect to domain I to form a different intermediate of the PET (Fig. 3a, b). A subsequent maturation step of states 2 and 3 is likely to involve the joining of domains III and VI and the formation of the PET

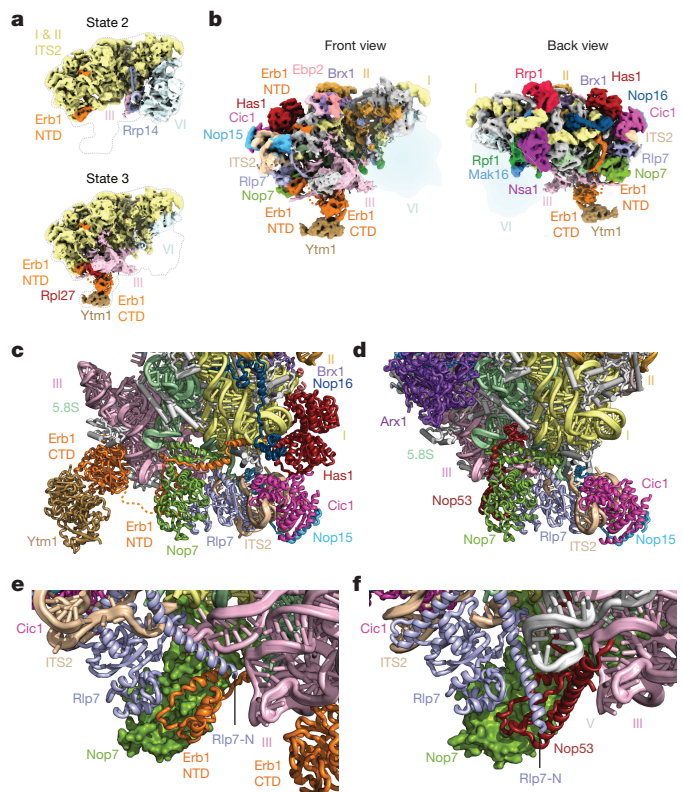


Figure 3 | Molecular mimicry by Erb1 prevents premature ITS2 processing. **a**, Cryo-EM maps of states 2 and 3 of the nucleolar pre-60S particle. State 2 contains domain VI but limited density for domain III, while state 3 contains domain III with a flexible domain VI. NTD, N-terminal domain. **b**, Two views of the state 3 cryo-EM map. **c, d**, Cartoon representation of the ITS2 region in state 3 of the nucleolar pre-60S particle (**c**) and the Nog2 particle (PDB 3JCT) (**d**). **e, f**, The N termini of Erb1 and Rlp7 in the nucleolar pre-60S particle prevent the binding of Nop53 (**e**), which binds to Nop7 in the Nog2 particle (**f**).

on the solvent-exposed side. This may be accompanied by the insertion of the Nog1 N terminus into the nascent PET and the replacement of Ssf1–Rrp15 by Rpl31 (Fig. 2c, d, 4).

Conformational changes, first by the initially flexible domain V–5S RNP together with the Nog1 GTPase domain, and subsequently by domain IV, would result in the overall conformation observed in the late-nucleolar Nog2 particle² (Fig. 2c, d, 4). The base-pairing between

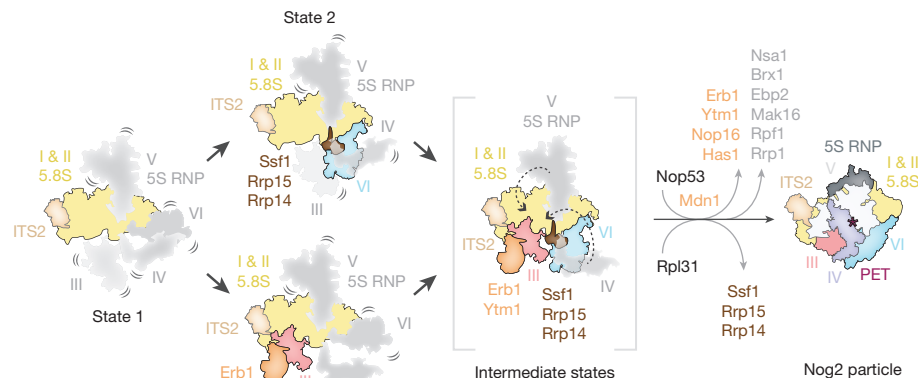


Figure 4 | Model of early nucleolar stages of large subunit assembly. Domains of 25S rRNA are represented as separate segments that are flexible (in grey) or stably incorporated (in colour). State 1 represents an intermediate in which domains I and II are ordered and subsequent folding of either domain III or VI will occur (state 3 and state 2, respectively) to partially form the polypeptide exit tunnel (PET). To mature into the Nog2 particle, the formation of the nascent PET must occur by Mdn1- and Rix7-dependent exchange and removal of assembly factors alongside stepwise folding of domain V and then domain IV.

domains I and V would be possible upon the dissociation of the Brx1–Ebp2 complex (Fig. 2e, f), while the release of Rrp1, Rpf1 and Mak16 is likely to be caused by the ATPase Rix7 acting on the proximal Nsa1 (ref. 15). In the ITS2 region, Mdn1-dependent removal of Erb1–Ytm1 would expose the binding site of Nop53 and may also trigger the exit of Nop16 and Has1 from the particle¹³.

While this manuscript was under review, complementary structural data on the nucleolar assembly of the large ribosomal subunit was published¹⁶. Differences in purification conditions resulted in the isolation of distinct states, which may represent assembly stages or breakdown products. Together with our data, these results enable visualization of early nucleolar pre-60S intermediates, from the formation of the solvent-exposed side of the particle to the incorporation of the DEAD-box helicase Spb4 at the subunit interface (Extended Data Fig. 7). Throughout the assembly, proteins such as Erb1, Brx1 and subsequently Spb1 prevent premature association of later maturation factors by steric hindrance (Extended Data Fig. 8). Additionally, flexible elements of Ebp2 and Erb1 reduce conformational freedom of the maturing particles.

Eukaryotic nucleolar 60S ribosome assembly is conceptually reminiscent of early prokaryotic 50S assembly intermediates in which different rRNA domains are assembled in a modular fashion¹⁷ (Fig. 4, Extended Data Fig. 7). However, our structures of nucleolar pre-60S particles highlight the high degree of control that is exerted to prevent the premature formation of inter-domain contacts of ribosomal RNA. They further illustrate how the reduction of conformational freedom and ordered sequence of assembly factors are enforced during assembly. These are overarching themes of the nucleolar stages for both the small and large ribosomal subunit assembly¹⁸.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 November 2017; accepted 23 February 2018.

Published online 5 March 2018.

- Konikkat, S. & Woolford, J. L. Jr. Principles of 60S ribosomal subunit assembly emerging from recent studies in yeast. *Biochem. J.* **474**, 195–214 (2017).
- Wu, S. *et al.* Diverse roles of assembly factors revealed by structures of late nuclear pre-60S ribosomes. *Nature* **534**, 133–137 (2016).
- Barrio-Garcia, C. *et al.* Architecture of the Rix1–Rea1 checkpoint machinery during pre-60S-ribosome remodeling. *Nat. Struct. Mol. Biol.* **23**, 37–44 (2016).
- Ma, C. *et al.* Structural snapshot of cytoplasmic pre-60S ribosomal particles bound by Nmd3, Lsg1, Tif6 and Reh1. *Nat. Struct. Mol. Biol.* **24**, 214–220 (2017).
- Chen, W., Xie, Z., Yang, F. & Ye, K. Stepwise assembly of the earliest precursors of large ribosomal subunits in yeast. *Nucleic Acids Res.* **45**, 6837–6847 (2017).
- Talkish, J. *et al.* Disruption of ribosome assembly in yeast blocks cotranscriptional pre-rRNA processing and affects the global hierarchy of ribosome biogenesis. *RNA* **22**, 852–866 (2016).

- Kos-Braun, I. C., Jung, I. & Koš, M. Tor1 and CK2 kinases control a switch between alternative ribosome biogenesis pathways in a growth-dependent manner. *PLoS Biol.* **15**, e2000245 (2017).
- Harnpicharnchai, P. *et al.* Composition and functional characterization of yeast 66S ribosome assembly intermediates. *Mol. Cell* **8**, 505–515 (2001).
- Clinton, C. & Gazda, H. T. in *Diamond–Blackfan Anemia*. (eds Adam, M. P. *et al.*) (University of Washington, 1993).
- Granneman, S., Petfalski, E. & Tollervey, D. A cluster of ribosome synthesis factors regulate pre-rRNA folding and 5.8S rRNA maturation by the Rat1 exonuclease. *EMBO J.* **30**, 4006–4019 (2011).
- Konikkat, S., Biedka, S. & Woolford, J. L., Jr. The assembly factor Erb1 functions in multiple remodeling events during 60S ribosomal subunit assembly in *S. cerevisiae*. *Nucleic Acids Res.* **45**, 4853–4865 (2017).
- Thoms, M. *et al.* The exosome is recruited to RNA substrates through specific adaptor proteins. *Cell* **162**, 1029–1038 (2015).
- Baßler, J. *et al.* The AAA-ATPase Rea1 drives removal of biogenesis factors during multiple stages of 60S ribosome assembly. *Mol. Cell* **38**, 712–721 (2010).
- Dembowski, J. A., Kuo, B. & Woolford, J. L. Jr. Has1 regulates consecutive maturation and processing steps for assembly of 60S ribosomal subunits. *Nucleic Acids Res.* **41**, 7889–7904 (2013).
- Kressler, D., Roser, D., Pertsch, B. & Hurt, E. The AAA ATPase Rix7 powers progression of ribosome biogenesis by stripping Nsa1 from pre-60S particles. *J. Cell Biol.* **181**, 935–944 (2008).
- Kater, L. *et al.* Visualizing the assembly pathway of nucleolar pre-60S ribosomes. *Cell* **171**, 1599–1610.e14 (2017).
- Davis, J. H. *et al.* Modular assembly of the bacterial large ribosomal subunit. *Cell* **167**, 1610–1622.e15 (2016).
- Barandun, J. *et al.* The complete structure of the small-subunit processome. *Nat. Struct. Mol. Biol.* **24**, 944–953 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Ebrahim and J. Sotiris for their support with data collection at the Evelyn Gruss Lipper Cryo-EM resource center, M. Tesic-Mark for analysis of mass spectrometry data, C. Cheng for help with the initial manual curation and analysis of the nucleolar pre-60S particles and members of the Walz laboratory for helpful discussions. L.M. is supported in part by NIH T32 GM115327-Tan. J.B. is supported by an EMBO long-term fellowship (ALTF 51-2014) and a Swiss National Science Foundation fellowship (155515). M.C.-M. is supported by a postgraduate scholarship from NSERC. S.K. is supported by the Robertson Foundation, the Irma T. Hirsch Trust, the Alexandrine and Alexander L. Sinsheimer Fund, the Rita Allen Foundation and an NIH New Innovator Award (1DP2GM123459). B.T.C. is supported by National Institute of Health Grant Nos. P41GM103314 and P41GM109824.

Author Contributions S.K. and Z.A.S. established purification conditions. Z.A.S., L.M. and S.K. determined the cryo-EM structure of the yeast nucleolar pre-60S particle. K.R.M., J.W. and B.T.C. processed and analysed DSS cross-linking data. Z.A.S., L.M., J.B., M.C.-M., M.H. and S.K. built the model. L.M. performed all RNA work, and Z.A.S., L.M., J.B., M.H., M.C.-M. and S.K. interpreted the results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.K. (klings@rockefeller.edu).

Reviewer Information Nature thanks A. Amunts, D. Tollervey and J. Woolford for their contribution to the peer review of this work.

METHODS

Purification of nucleolar pre-60S particles. Nucleolar pre-60S particles were purified from a *Saccharomyces cerevisiae* BY4741 strain containing a TEV protease-cleavable C-terminal GFP tag on Nsa1 (Nsa1–40aaLinker–TEV–GFP) and a C-terminal 5× beta-catenin 3C protease-cleavable tag on Nop2 (Nop2–40aaLinker–3C–Bc5) for endogenous expression. Cultures were grown in full synthetic drop-out (SD) medium containing 2% raffinose (w/v) at 30 °C to an optical density of 0.8–1, before addition of 2% galactose (w/v) for 16 h, reaching saturation (optical density 5–6). Cells were then harvested by centrifugation at 3,000g for 10 min at 4 °C. The cell pellet was washed with ice-cold ddH₂O twice, followed by a wash with ddH₂O containing protease inhibitors (E64, pepstatin and PMSF). Washed cells were immediately flash frozen in liquid nitrogen and lysed by four cycles of cryogenic grinding using a Retsch Planetary Ball Mill PM100.

The freshly ground yeast powder was resuspended by vortexing in buffer A (50 mM Tris–HCl, pH 7.6 (20 °C), 150 mM NaCl, 1 mM EDTA, 1 mM DTT, 0.1% Triton-X100, PMSF, pepstatin, E-64). The insoluble fraction was removed by centrifugation at 4 °C, 40,000g for 30 min. The supernatant was subsequently incubated with anti-GFP nanobody beads (Chromotek) for 3 h at 4 °C, with agitation. The beads were washed four times in ice-cold buffer A before the bound proteins were eluted using TEV-protease cleavage (1 h, 4 °C). The eluate was then incubated with NHS-sepharose beads (Sigma) coupled with anti-beta-catenin nanobody¹⁹ in buffer B (50 mM Tris–HCl pH 7.6 (20 °C), 150 mM NaCl, 1 mM EDTA, 1 mM DTT) for 1 h at 4 °C with agitation. For electron microscopy sample preparation, the anti-beta catenin beads were washed once with buffer B. Cleavage by 3C protease for 1 h at 4 °C released the Nsa1–Nop2 containing nucleolar pre-60S particles. For protein–protein cross-linking analysis the eluate from GFP-nanobody beads was incubated with beta catenin-nanobody beads in buffer C (50 mM HEPES–NaOH pH 7.6 (4 °C), 150 mM NaCl, 1 mM EDTA) and eluted in the same buffer by 3C-protease cleavage. The eluate typically measured an absorbance at 260 nm (A_{260}) of 2.4–4.5 milli absorbance units (mAU) (Nanodrop 2000, Thermo Scientific) (Extended Data Fig. 1).

Cryo-EM sample and grid preparation. Cryo-EM grids were prepared on four different occasions for the four datasets obtained (ds1–ds4). The nucleolar pre-60S particle eluate in sample buffer B (above) was left as is (ds1 only), or supplemented with 0.1% Triton X-100 and 5 mM MgCl₂ (final concentration, ds2–ds4). Copper grids of 400 mesh with lacey carbon and an ultra-thin carbon support film were used (Ted Pella, product no. 01824) for data collection. A volume of 3–4 µl nucleolar pre-60S particle sample (absorbance at 260 nm of 2.5 mAU) was applied onto glow-discharged grids and plunged into liquid ethane using a Vitrobot Mark IV robot (FEI Company) (100% humidity, blot force of 0 and blot time 3.5–4 s).

Cryo-EM data collection and image processing. A total of 14,201 micrographs were obtained over four data collections (ds1–ds4) on a Titan Krios (FEI Company) at 300 kV with a K2 Summit detector (Gatan). SerialEM²⁰ was employed for data acquisition using a defocus range of 1.0–3.5 µm with a pixel size of 1.3 Å. Super-resolution movies with 32 frames were collected using a total dose of 10 electrons per pixel per second with an exposure time of 8 s and a total dose of 47 electrons per Å² (Table 1).

Upon data collection, the movies were gain corrected, dose weighted and aligned with Motioncor2 (ref. 21), and the contrast transfer function (CTF) was estimated using CTFFIND 4.1.5 (ref. 22). Relion 2.1 (ref. 23) was used for all subsequent particle picking, classifications and refinements. Corrected and aligned micrographs were first subjected to autopicking in Relion, resulting in a total of 1,653,290 selected particles from all four datasets. After manual inspection of all micrographs, particles were extracted with a box size of 480 pixels (2×-binned to 240 pixels), and 2D-classified separately for each individual dataset. After 2D classification, bad classes were removed, and the selected particles of each dataset were 3D-classified into four classes using an initial 3D model obtained from cryoSPARC²⁴, low-pass filtered to 60 Å. The best one-to-two classes from each 3D classification were selected and their particles were re-extracted with a box size of 480 pixels (un-binned). A combined total of 514,746 particles was finally used for 3D auto-refinement and post-processing with a solvent mask around a 'core' containing domains I and II of the 25S rRNA, resulting in an overall resolution of 3.4 Å (Extended Data Figs 2, 3a). Since the more flexible domains III–VI were not visible in this state owing to averaging of different populations, a subsequent round of 3D classification without image alignment of these particles into six classes was used to obtain the two classes that contained state 2 (39%) and states 1 and 3 (41%). A refinement of the class containing state 2 (201,114 particles) was performed using a mask to include the additional visible densities, comprising of domain VI and its associated proteins, to obtain a final map at 3.7 Å (Extended Data Figs 2, 3b). By performing an additional round of focused 3D classification without image alignment on this class with a mask around Mak11 and the neighbouring segment of domain V of the 25S rRNA, a subset of particles (36%) emerged with improved

density. This was refined to provide a more continuous map of Mak11 and domain V (state 2A) (Extended Data Figs 2, 3c). The remaining 64% of particles did not reveal additional features beyond those seen in state 2.

Owing to particle heterogeneity in the initial class containing states 1 and 3 (41%, 211,534 particles), a focused classification without image alignment of that class of particles was performed using a mask around the additional density containing domain III of the 25S rRNA, Erb1-CTD (WD40) and Ytm1 (WD40). The class from this round of classification containing density for domain III (31,419 particles) was refined with a mask around the entirety of state 3, to obtain a final map at 4.6 Å resolution (Extended Data Figs 2, 3d). The highest-resolution class of particles from this classification, which contained no density for domain III (126,824 particles), was refined to obtain the 4.3 Å map of state 1. The map of state 1 closely resembles the higher-resolution core map (correlation coefficient = 0.96). The local resolutions of the maps were calculated using Resmap²⁵ (Extended Data Fig. 3).

Model building and refinement. By using the structure of the late nucleolar pre-60S particle—the Nog2 particle (PDB 3JCT)²—as reference, common assembly factors and ribosomal proteins were manually located and fitted into the density, using Cic1 and Rpl7 as hallmark anchors. Protein identification and tracing were aided by cross-linking and mass spectrometry analyses (described below). New assembly factors Mak16, Rrp1, Nop16, Erb1-NTD, Rrp14, Rrp15 and Ebp2 and segments of rRNA were modelled *de novo*. Previously determined crystal structures of assembly factors Nsa1 (PDB 5SUI)²⁶, Ytm1 (PDB 5CXB) and Erb1-CTD (PDB 4U7A)²⁷ were docked and manually adjusted. Has1, Ssf1, Brx1, Rpf1 and Mak11 were initially docked from Phyre2 models²⁸, and manually built to fit the density. Model building was performed with Coot²⁹. An annotated list of individual protein IDs, reference models and corresponding maps used for building can be found in Extended Data Table 1. The model was refined against a half-map1 from the overall 3.7 Å map of state 2 in PHENIX with phenix.real_space_refine using secondary structure restraints for proteins and RNAs³⁰. Refinement and model statistics can be found in Table 1.

Map and model visualization. All map and model analyses and illustrations were made using Chimera³¹ and PyMOL Molecular Graphics System, v.1.8 (Schrödinger). Density map visualization for certain figures was also performed on UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics and the University of California, San Francisco (supported by NIGMS P41-GM103311).

RNA extraction and northern blotting. The *S. cerevisiae* nucleolar pre-60S particle (Nsa1–Nop2 particle) was purified as described in 'Purification of nucleolar pre-60S particles' and RNA was extracted from the final 3C-protease elution with 1 ml TRIzol (Life Technologies) according to the manufacturer's instructions. Isolated nucleolar pre-60S RNA (1.0 µg) was separated on a denaturing 1.2% formaldehyde–agarose gel (SeaKem LE, Lonza) or on a denaturing 10% urea–PAGE (Fisher, Amresco) for the 5S rRNA northern blot. After staining the gel in 1× SYBR Green II (Lonza) ddH₂O solution (pH 7.5) for 30 min, RNA species were visualized with a Gel Doc EZ Imager (Bio-Rad) (Extended Data Fig. 1d, e) and then transferred onto a cationized nylon membrane (Zeta-Probe GT, Bio-Rad) using downward capillary transfer for the agarose gel and a Trans-Blot SD semi-dry transfer cell (Bio-Rad) for the urea–PAGE gel. RNA was cross-linked to the membrane for northern blot analysis by UV irradiation at 254 nm with a total exposure of 120 mJ/cm² in a UV Stratalinker 2400 (Stratagene). Cross-linked membranes were incubated with hybridization buffer (750 mM NaCl, 75 mM trisodium citrate, 1% (w/v) SDS, 10% (w/v) dextran sulphate, 25% (v/v) formamide) at 65 °C for 30 min before addition of γ -³²P-end-labelled DNA oligonucleotide probes. Oligonucleotide probe sequences were as follows: 25S, TTTCACTCTCTTTTCAAAGTTCTTTTCATCT; ITS1 3' end, TT AATATTTTAAATTTCCAG; ITS2 C2 site, TGGTAAACCTAAACGACCGT; 3'-ETS 5' end, CCACTTAGAAAGAAATAAAAA; and 5S, CTACTCG GTCAGGCTC.

Probes were hybridized for 1 h at 65 °C and then overnight at 37 °C. Membranes were washed once with wash buffer 1 (300 mM NaCl, 30 mM trisodium citrate, 1% (w/v) SDS) and once with wash buffer 2 (30 mM NaCl, 3 mM trisodium citrate, 1% (w/v) SDS) for 20 min each at 45 °C. Radioactive signal was detected by exposure of the washed membranes to a storage phosphor screen which was scanned with a Typhoon 9400 variable-mode imager (GE Healthcare). For northern blot source data, see Supplementary Fig. 1.

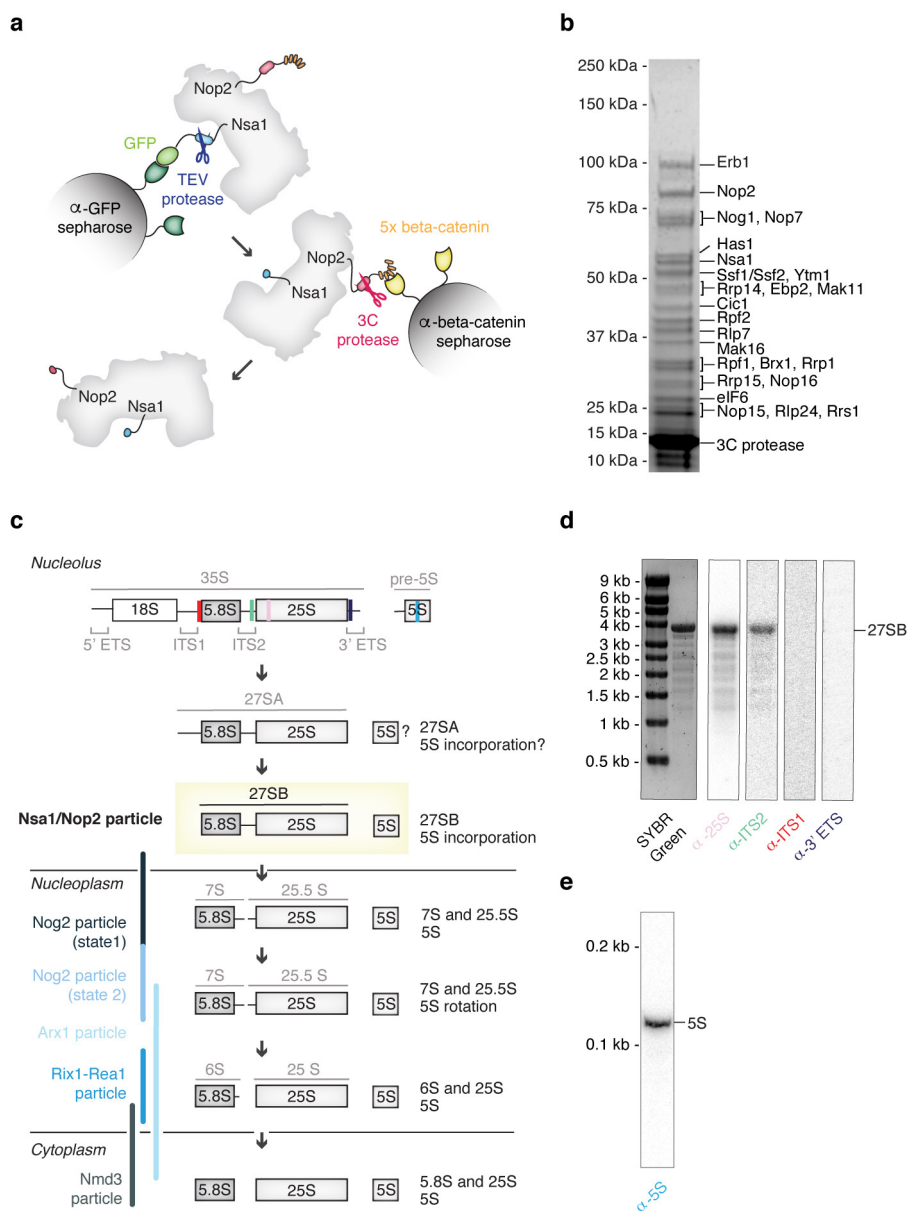
DSS cross-linking sample preparation and mass spectrometry analysis. The tandem-affinity purified nucleolar pre-60S particles (Nsa1–Nog2 particle), eluted off anti-beta catenin nanobody beads (in 50 mM HEPES–NaOH pH 7.6 (4 °C), 150 mM NaCl, 1 mM EDTA, 1 mM DTT) at an absorbance of 1.0 at 260 nm (Nanodrop 2000, Thermo Scientific), were pooled (total volume, 300 µl) and split into three 100-µl cross-linking reaction aliquots.

Disuccinimidylsuberate (DSS; 25 mM in DMSO, Creative Molecules) was added to each aliquot to yield a final DSS concentration of 2.0 mM and samples were cross-linked for 30 min at 25 °C with 450 r.p.m. constant mixing. The reactions were quenched with 50 mM ammonium bicarbonate (final concentration) and precipitated by adding methanol (Alfa Aesar, LC-MS grade) to a final concentration of 90% followed by incubation at −80 °C overnight. Precipitated cross-linked nucleolar pre-60S particles were combined into one tube by repeated centrifugation at 21,000g, 4 °C for 30 min. The resulting pellet was washed three times with 1 ml ice-cold 90% methanol, air-dried and finally resuspended in 50 µl 1 × NuPAGE LDS buffer (Thermo Fisher Scientific).

DSS cross-linked samples were processed as in ref. 18, and as described below. DSS cross-linked nucleolar pre-60S particles in LDS buffer were reduced with 25 mM DTT, alkylated with 100 mM 2-chloroacetamide, separated by SDS-PAGE in three lanes of a 3–8% tris-acetate gel (NuPAGE, Thermo Fisher Scientific), and stained with Coomassie blue. The gel region corresponding to cross-linked complexes was sliced and digested overnight with trypsin to generate cross-linked peptides. After digestion, the peptide mixture was acidified and extracted from the gel as previously described^{32,33}. Peptides were fractionated offline by high-pH reverse-phase chromatography, loaded onto an EASY-Spray column (Thermo Fisher Scientific ES800: 15 cm × 75 µm ID, PepMap C18, 3 µm) via an EASY-nLC 1000, and gradient-eluted for online electrospray ionization–mass spectrometry (ESI-MS) and tandem mass spectrometry (MS/MS) analyses with a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific). MS/MS analyses of the top 8 precursors in each full scan used the following parameters: resolution, 17,500 (at 200 Th); AGC target, 2×10^5 ; maximum injection time, 800 ms; isolation width, 1.4 m/z; normalized collision energy, 24%; charge, 3–7; intensity threshold, 2.5×10^3 ; peptide match, off; dynamic exclusion tolerance, 1,500 mmu. Cross-linked peptides were identified from mass spectra by pLink³⁴. All spectra reported here were manually verified as described previously³².

Data availability. Cryo-EM density maps for the yeast nucleolar pre-60S particle states 2 and 3 have been deposited in the EM Data Bank with accession codes EMD-7324 and EMD-7445, respectively. Atomic coordinates for the yeast nucleolar pre-60S particle states 2 and 3 have been deposited in the Protein Data Bank under accession codes 6C0F and 6CB1, respectively. A PyMOL session for the analysis of the yeast nucleolar pre-60S particle in state 2 is available in Supplementary Dataset 2.

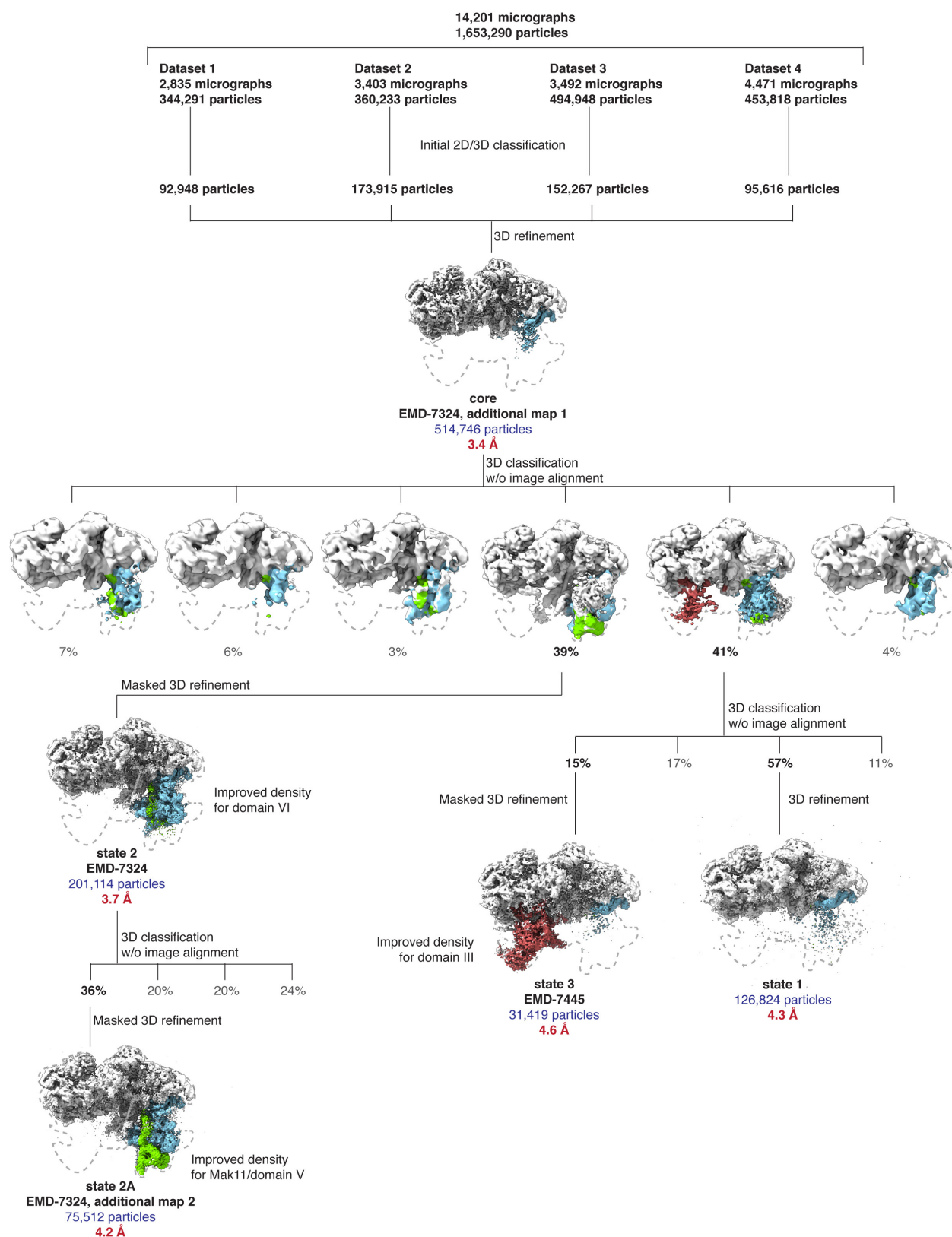
19. Braun, M. B. *et al.* Peptides in headlock—a novel high-affinity and versatile peptide-binding nanobody for proteomics and microscopy. *Sci. Rep.* **6**, 19211 (2016).
20. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
21. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
22. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
23. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, 19 (2016).
24. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
25. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
26. Lo, Y.-H., Romes, E. M., Pillon, M. C., Sobhany, M. & Stanley, R. E. Structural analysis reveals features of ribosome assembly Factor Nsa1/WDR74 important for localization and interaction with Rix7/NVL2. *Structure* **25**, 762–772.e4 (2017).
27. Wegrecki, M., Rodríguez-Galán, O., de la Cruz, J. & Bravo, J. The structure of Erb1-Ytm1 complex reveals the functional importance of a high-affinity binding between two β-propellers during the assembly of large ribosomal subunits in eukaryotes. *Nucleic Acids Res.* **43**, 11017–11030 (2015).
28. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
29. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
30. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
31. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
32. Shi, Y. *et al.* Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol. Cell. Proteomics* **13**, 2927–2943 (2014).
33. Shi, Y. *et al.* A strategy for dissecting the architectures of native macromolecular assemblies. *Nat. Methods* **12**, 1135–1138 (2015).
34. Yang, B. *et al.* Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
35. Bradatsch, B. *et al.* Structure of the pre-60S ribosomal subunit with nuclear export factor Arx1 bound at the exit tunnel. *Nat. Struct. Mol. Biol.* **19**, 1234–1241 (2012).



Extended Data Figure 1 | Purification of tagged Nsa1–Nop2 nucleolar pre-60S particles and analysis of RNA components.

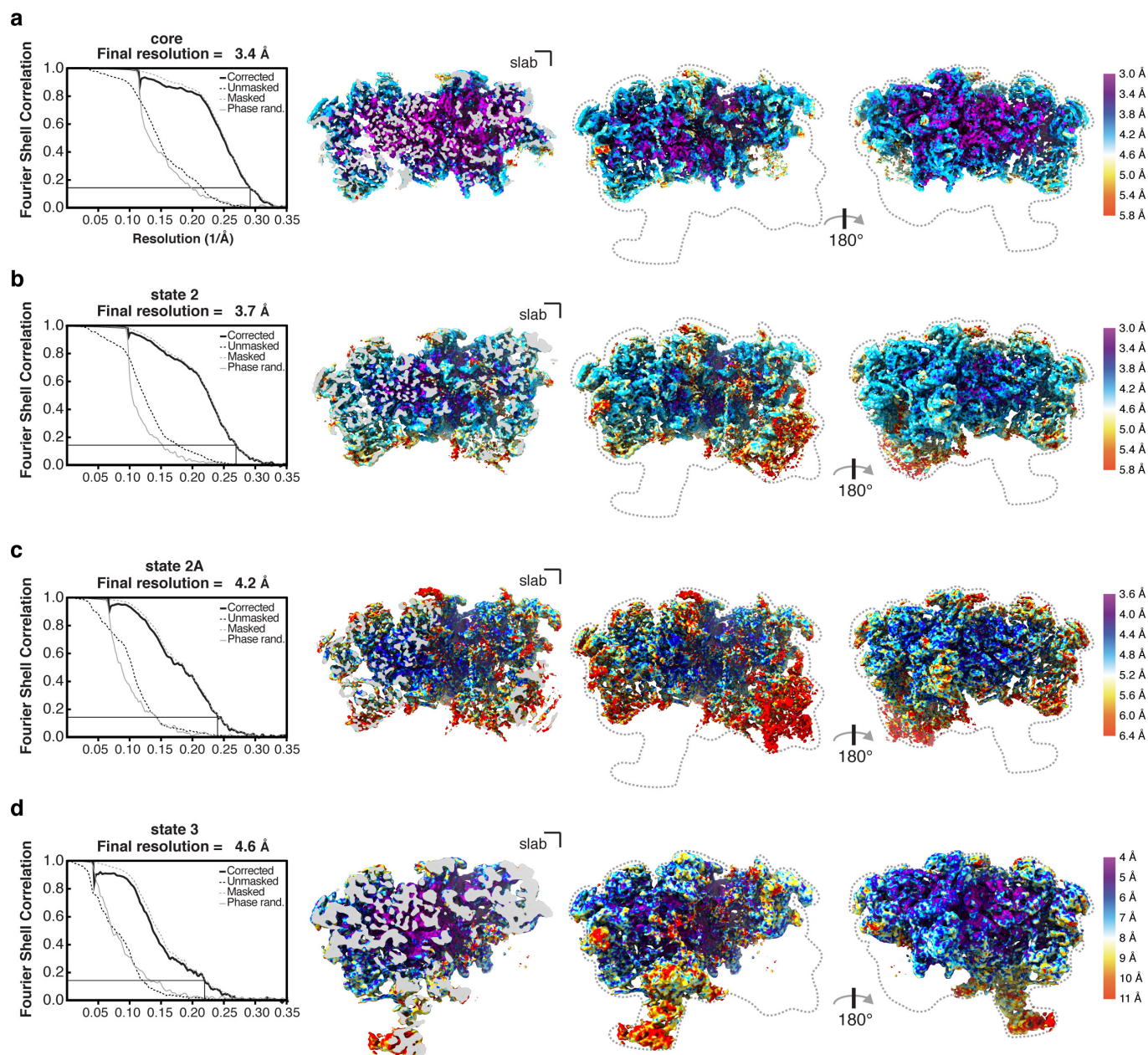
a, Schematic of tandem-affinity purification of the nucleolar pre-60S particle with tagged proteins Nsa1 and Nop2. **b**, Coomassie-blue stained SDS–PAGE of pre-60S particles purified as in **a**. Protein labels are based on in-solution mass spectrometry analysis of purified pre-60S particles and the approximate molecular weight. **c**, Schematic processing of the

large ribosomal subunit rRNAs in yeast. The locations of the previously published pre-60S particles (the Nog2 particles², the Arx1 particle³⁵, the Rix1–Rea1 particle³ and the Nmd3 particle⁴) are represented by blue bars. Binding sites of northern blot probes are indicated on the 35S and pre-5S transcript. **d**, **e**, Pre-rRNA was visualized on an agarose gel and stained using SYBR Green II. Northern blot analysis was performed for the 25S, ITS2, ITS1 and 3' ETS rRNAs (**d**) and for the 5S rRNA (**e**).



Extended Data Figure 2 | Cryo-EM data-processing workflow. Four data collections were performed, resulting in 14,201 micrographs. These micrographs were aligned using MotionCor2 (ref. 21) with dose weighting, and imported into Relion2.1 (ref. 23) for further processing. Autopicking followed by manual cleaning, 2D and 3D classification produced a total

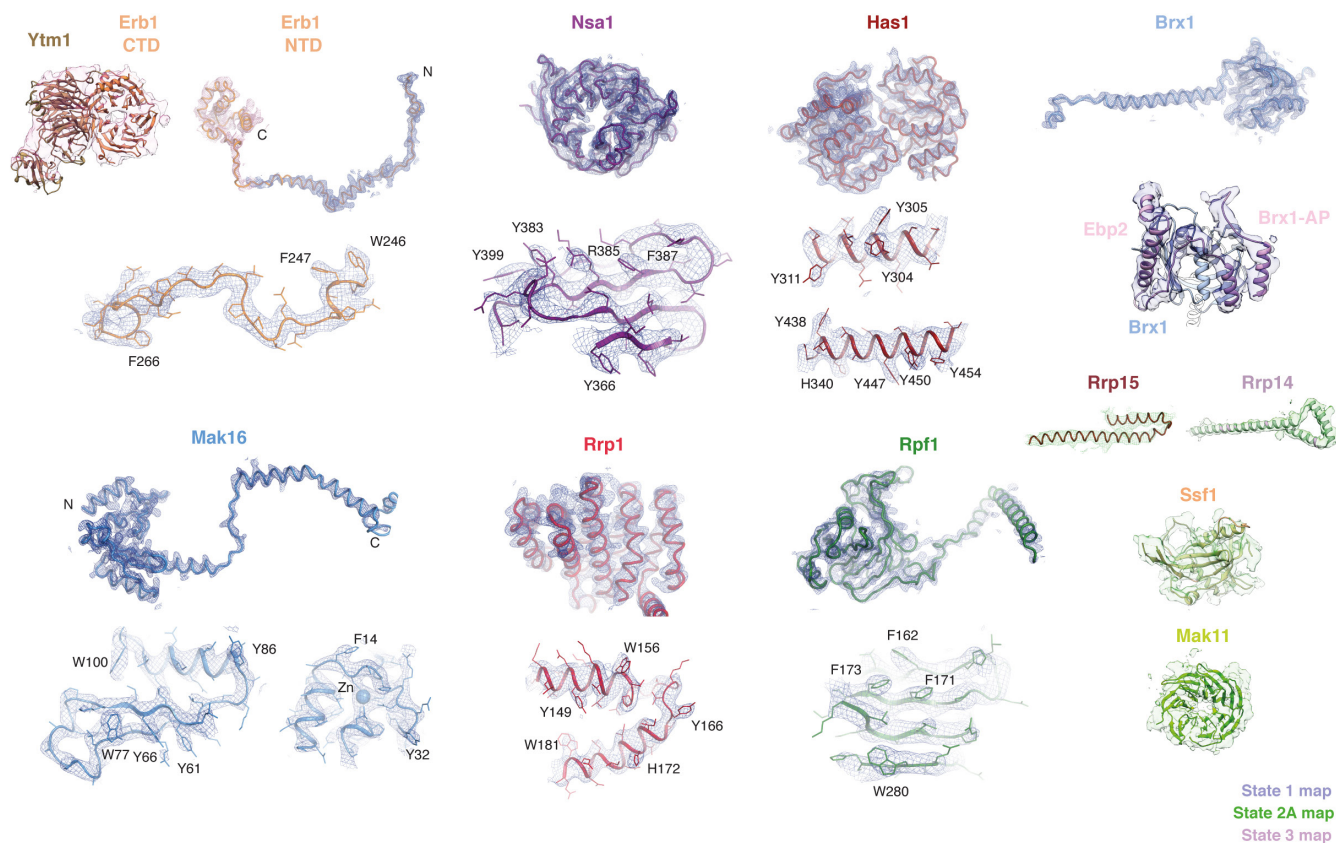
of 514,746 'good' particles. These particles were refined to produce the core map. Further 3D classification without and with alignment was used to obtain the state 1, state 2, state 2A and state 3 maps. Density regions corresponding to domain III (red), domain VI (blue) and Mak11 (green) are highlighted.



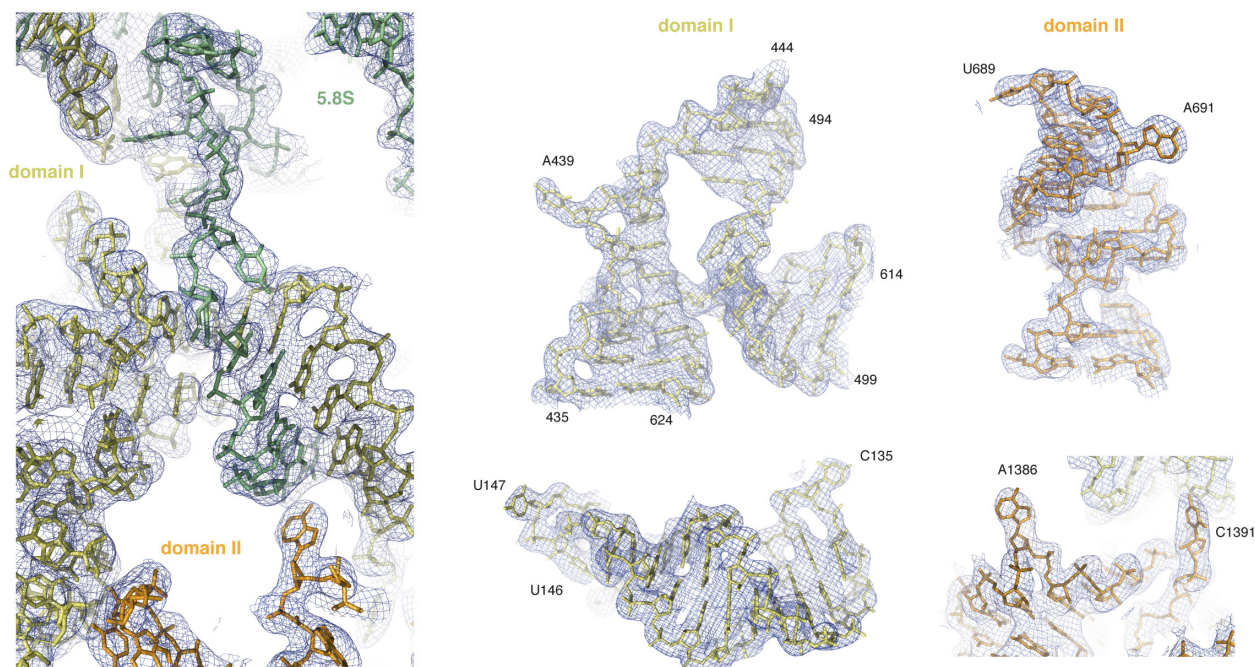
Extended Data Figure 3 | Overall and local resolution estimates for core, state 2 and state 3 cryo-EM maps. a–d, Overall and local resolution of core map at 3.4 Å (a), state 2 map at 3.7 Å (b), state 2A map with additional density for Mak11 at 4.2 Å (c) and state 3 map at 4.6 Å (d). Fourier Shell Correlation (FSC) curves for the unmasked (dashed black

line), phase-randomized (solid grey line), masked (dashed grey line) and the corrected map (solid black line) are also shown. A thin black line indicates an FSC value of 0.143. A clipped view is shown next to two views of the obtained cryo-EM maps. The density volumes are coloured according to local resolution using Resmap²⁵.

a



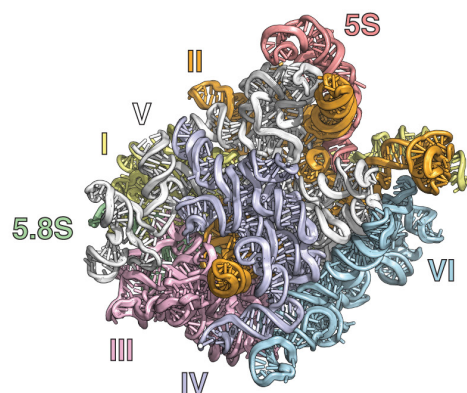
b



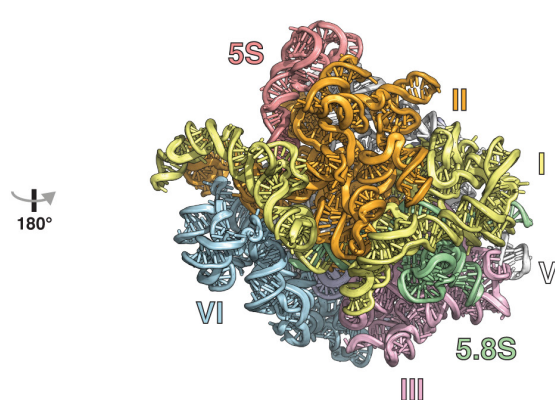
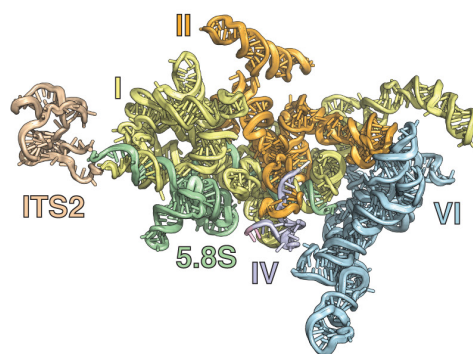
Extended Data Figure 4 | Cryo-EM density fit of selected nucleolar pre-60S ribosomal subunit proteins and RNA models. a, Near-atomic models of assembly factors and their cryo-EM density. **b,** Selected regions

of the 25S rRNA and 5.8S rRNA models and their cryo-EM density. Images generated in PyMOL or Chimera.

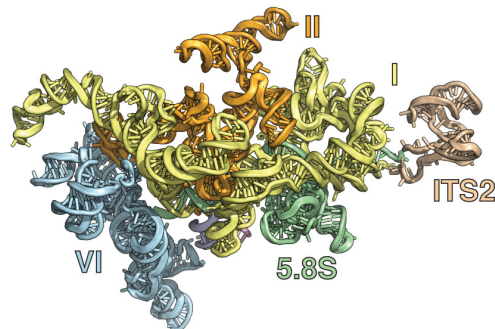
crown view



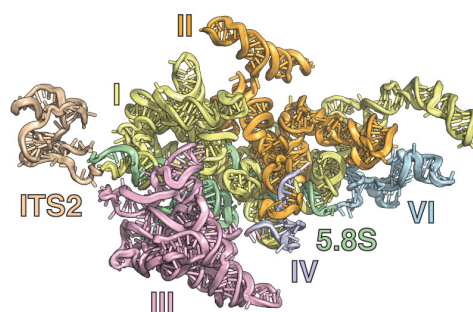
back view

mature
60S subunit

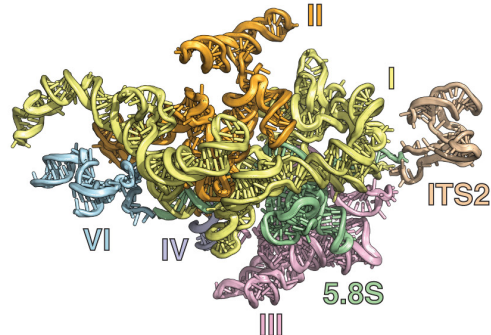
180°



state 2



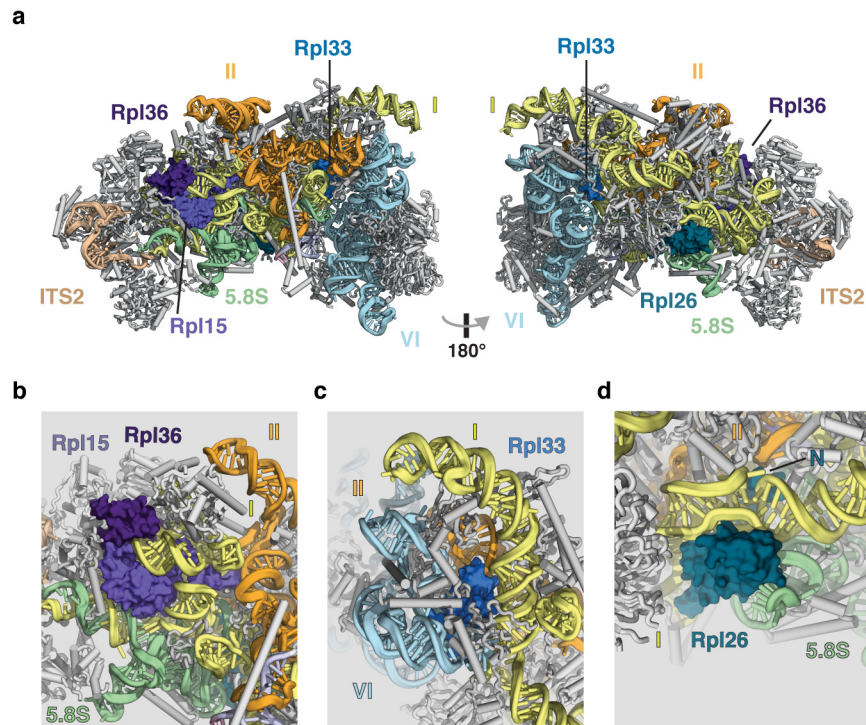
180°



state 3

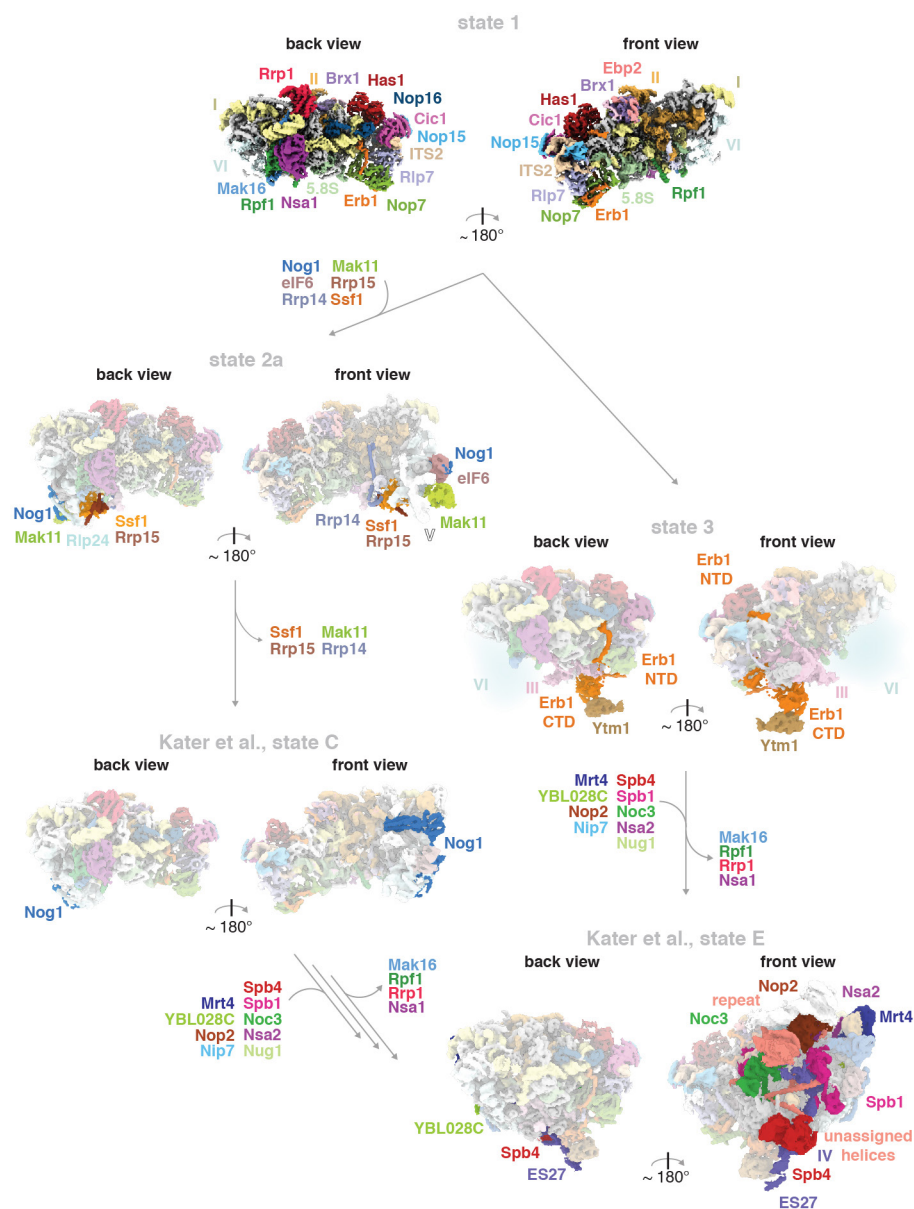
Extended Data Figure 5 | rRNA domains of state 2, state 3 and the mature 60S ribosomal subunit. The 5.8S rRNA, the 5S rRNA and the

domains of the 25S rRNA are colour-coded and displayed in the crown and back view for each structure.



Extended Data Figure 6 | Ribosomal proteins associated with Diamond-Blackfan anaemia are positioned at rRNA domain junctions in the nucleolar pre-60S particle. **a**, Two views of the nucleolar pre-60S particle state 2 model, with Diamond-Blackfan anaemia-associated ribosomal proteins shown in surface representation. **b**, Rpl15 is located

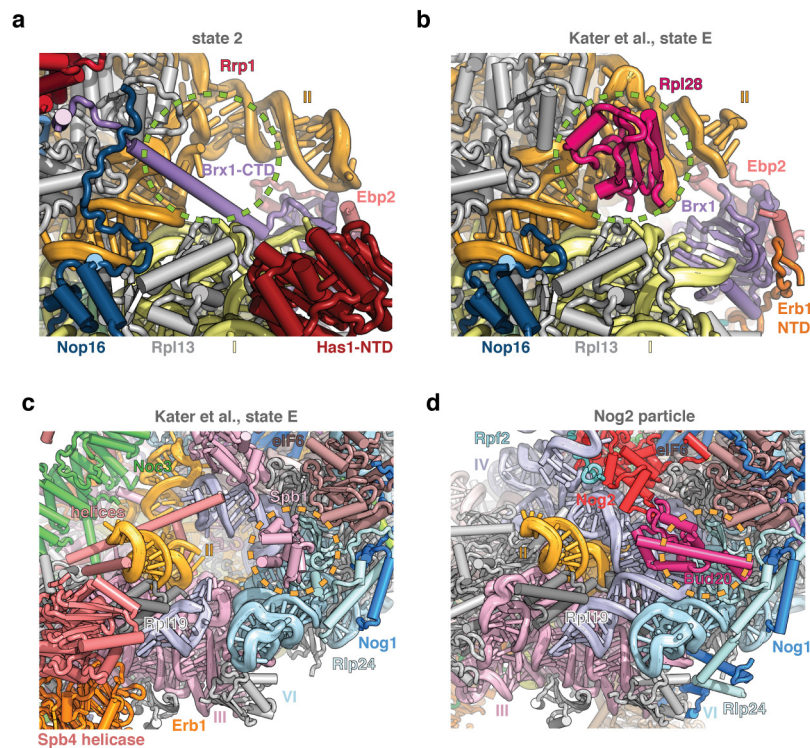
between the 5.8S-domain I duplex and a domain I-domain II interface. **c**, Rpl33 (Rpl35 in *Homo sapiens*) binds at the junction of domains I, II and VI in the 25S rRNA. **d**, Rpl26 associates with the domain I-5.8S rRNA interface and additionally inserts its N terminus (N) between domain I and domain II.



Extended Data Figure 7 | Intermediates of nucleolar pre-60S assembly.

The structural data presented in this paper (states 1, 2, and 3) are complemented by recent data on pre-60S assembly: states C (EMD-3893) and E (EMD-3891)¹⁶. State 1 is highly similar to state A (EMD-3888)¹⁶. States 2 and 2A correspond closely to state B (EMD-3889)¹⁶, but state B

lacks the Ssf1–Rrp15–Rrp14 module and Mak11. Built assembly factors that become ordered or leave in subsequent particles are indicated with arrows. Two possible pathways are shown that result in the final incorporation of the DEAD-box helicase Spb4 (previously unidentified¹⁶).



Extended Data Figure 8 | Steric hindrance during nucleolar pre-60S assembly. **a, b**, Comparative views of state 2 (**a**) and state E (PDB 6ELZ)¹⁶ (**b**), highlighting that the binding of the Brx1 CTD to Rrp1 prevents premature

incorporation of Rpl28. **c, d**, Comparative views of state E (**c**) and the Nog2 particle (PDB 3JCT) (**d**), highlighting that the presence of Spb1 prevents the binding of Bud20.

Extended Data Table 1 | Molecular models of the nucleolar pre-60S ribosomal subunit

Subgroup	Chain ID	SegID	Molecule name	Total residues or bases	Modelled (residue range)	Initial PDB template	Present in...
RNA	1	L1	25S	3,396	atomic (1,676 bases)	3JCT	All states
	2	L2	5.8S	158	atomic (158 bases)	3JCT	All states
	6	L6	ITS2	232	atomic (87 bases)	3JCT	All states
Ribosomal proteins	C	LC	Rpl4A_uL4	362	atomic (2-56, 89-347)	3JCT	All states
	E	LE	Rpl6A_eL6	176	atomic (7-176)	3JCT	All states
	e	SE	Rpl32_eL32	130	atomic (7-36, 47-130)	3JCT	All states
	F	LF	Rpl7A_uL30	244	atomic (3-244)	3JCT	All states
	f	SF	Rpl33A_eL33	107	atomic (2-107)	3JCT	All states
	G	LG	Rpl8A_eL8	256	atomic (53-239)	3JCT	All states
	h	SH	Rpl35A_uL29	120	atomic (2-120)	3JCT	All states
	i	SI	Rpl36A_eL36	100	atomic (17-100)	3JCT	All states
	L	LL	Rpl13A_eL13	199	atomic (22-127)	3JCT	All states
	M	LM	Rpl14A_eL14	138	atomic (11-138)	3JCT	All states
	N	LN	Rpl15A_eL15	204	atomic (2-68, 96-204)	3JCT	All states
	O	LO	Rpl16A_uL13	199	atomic (3-59, 73-199)	3JCT	All states
	Q	LQ	Rpl18A_eL18	186	atomic (15-146)	4v88, chain BQ	All states
	S	LS	Rpl20A_eL20	172	atomic (2-172)	3JCT	All states
	B	LB	Rpl3_uL3	387	atomic (17-224, 270-385)	4v88, chain DB	State 2
	P	LP	Rpl17A_uL22	184	atomic (10-64, 80-126, 140-161)	3JCT	All states
	V	LV	Rpl23A_uL14	137	atomic (16-137)	3JCT	State 2
	Y	LY	Rpl26A_uL24	127	atomic (2-127)	3JCT	All states
	j	SJ	Rpl37A_eL37	88	atomic (14-85, Zn)	3JCT	All states
	Z	LZ	Rpl27A_eL27	136	side-chain trimmed (2-136)	3JCT	State 3
	k	SK	Rpl38_eL38	78	side-chain trimmed (2-78)	3JCT	State 3
	g	SG	Rpl34A_eL34	121	side-chain trimmed (2-102)	3JCT	State 3
	c	SC	Rpl30_eL30	105	side-chain trimmed (9-105)	3JCT	State 3
	X	LX	Rpl25_uL23	142	side-chain trimmed (2-142)	3JCT	State 3
Assembly factors common to Nog2 particle (3JCT)	K	LK	Cic1	376	atomic (31-51, 64-302)	3JCT	All states
	n	SN	Nop7	605	atomic (13-43, 61-267, 351-396, 404-460)	3JCT	All states
	o	SO	Nop15	220	atomic (88-220)	3JCT	All states
	t	ST	Rlp7	322	atomic (54-105, 127-322)	3JCT	All states
	t	ST	Rlp7 (NTD)	322	poly-alanine (20-53)	De novo	State 3
	u	SU	Rlp24	199	atomic (2-130, Zn)	3JCT	State 2
	y	SY	Tif6	245	atomic (1-226)	3JCT	State 2
New assembly factors	W	LW	Nog1	647	atomic (373-470)	3JCT	State 2
	A	LA	Nsa1	463	atomic (1-78, 101-416)	5SUI	All states
	p	SP	Has1	505	atomic (42-252, 264-489)	Phyre model based on 2V1X	All states
	b	SB	Brx1	291	atomic (31-122, 132-164, 174-192, 211-290), poly-alanine (123-131, 165-173)	Model based on 5WLC, chain SM	All states
	m	SM	Ebp2	427	poly-alanine (196-269)	De novo	All states
	z	SZ	Rrp1	278	atomic (1-186, 197-253)	De novo	All states
	D	LD	Mak16	306	atomic (2-191, Zn)	De novo	All states
	l	LI	Rpf1	295	atomic (8-295)	Phyre model based on 5JPQ, chain c	All states
	s	SS	Erb1-NTD	807	atomic (239-298, 372-395), poly-alanine (299-371)	De novo	All states
	s	SS	Erb1-CTD	807	side-chain trimmed crystal structure (416-426, 428-534, 571-807)	4U7A	State 3
	v	SV	Ssf1	453	atomic (23-214, 324-356)	Phyre model based on 4XV9	State 2
	q	SQ	Mak11	468	poly-alanine (WD40, 285 residues)	Phyre model based on 3DM0	States 2 and 2A
	w	SW	Rrp15	250	atomic (174-243)	De novo	State 2
	d	SD	Ytm1	460	poly-alanine (<i>C. thermophilum</i> , 465 residues)	5CXB	State 3
	7	S7	Nop16	231	atomic (1-83, 156-228)	De novo	All states
	8	S8	Rrp14	434	atomic (296-393)	De novo	State 2
	x	SX	Brx1-associated peptide	unknown	poly-alanine (162-189)	De novo	All states

Individual protein chains are listed with their initial PDB template (or built *de novo*) and the nucleolar pre-60S particle state(s) in which they are present.

Structural insights into the voltage and phospholipid activation of the mammalian TPC1 channel

Ji She^{1,2,*}, Jiangtao Guo^{3,*}, Qingfeng Chen^{1,2,4,*}, Weizhong Zeng^{1,2,4}, Youxing Jiang^{1,2,4} & Xiao-chun Bai^{2,5}

The organellar two-pore channel (TPC) functions as a homodimer, in which each subunit contains two homologous *Shaker*-like six-transmembrane (6-TM)-domain repeats¹. TPCs belong to the voltage-gated ion channel superfamily² and are ubiquitously expressed in animals and plants^{3,4}. Mammalian TPC1 and TPC2 are localized at the endolysosomal membrane, and have critical roles in regulating the physiological functions of these acidic organelles^{5–7}. Here we present electron cryo-microscopy structures of mouse TPC1 (MmTPC1)—a voltage-dependent, phosphatidylinositol 3,5-bisphosphate (PtdIns(3,5)P₂)-activated Na⁺-selective channel—in both the apo closed state and ligand-bound open state. Combined with functional analysis, these structures provide comprehensive structural insights into the selectivity and gating mechanisms of mammalian TPC channels. The channel has a coin-slot-shaped ion pathway in the filter that defines the selectivity of mammalian TPCs. Only the voltage-sensing domain from the second 6-TM domain confers voltage dependence on MmTPC1. Endolysosome-specific PtdIns(3,5)P₂ binds to the first 6-TM domain and activates the channel under conditions of depolarizing membrane potential. Structural comparisons between the apo and PtdIns(3,5)P₂-bound structures show the interplay between voltage and ligand in channel activation. These MmTPC1 structures reveal lipid binding and regulation in a 6-TM voltage-gated channel, which is of interest in light of the emerging recognition of the importance of phosphoinositide regulation of ion channels.

TPC1 and TPC2 represent two major subfamilies of mammalian TPC channels and their functions are associated with various physiological processes, including hair pigmentation^{8–10}, autophagy regulation^{11,12}, blood vessel formation¹³, acrosome reaction in sperm¹⁴, mTOR-dependent nutrient sensing¹⁵, lipid metabolism¹⁶ and Ebola virus infection¹⁷, to name a few. Mammalian TPCs were initially suggested to mediate nicotinic acid adenine dinucleotide phosphate (NAADP)-dependent calcium release from endolysosomes^{18–20}. However, several recent studies have demonstrated that mammalian TPCs are Na⁺-selective channels activated by endolysosome-specific PtdIns(3,5)P₂ rather than NAADP^{15,21}. The dual regulation of TPC2 by both PtdIns(3,5)P₂ and NAADP has also been reported²². Distinct from TPC2, mammalian TPC1 activation is voltage-dependent, conferring electrical excitability to the endolysosome^{23,24}. The atomic structure of a plant TPC1 from *Arabidopsis thaliana* (AtTPC1) was recently determined by X-ray crystallography, revealing the overall architecture of the TPC family^{25,26}. However, mammalian TPCs share low sequence identity with their plant counterparts (Extended Data Fig. 1) and exhibit different gating and selectivity properties. Here we present the structural and functional analysis of MmTPC1.

When overexpressed in HEK293 cells, some MmTPC1 channels are trafficked to the plasma membrane, enabling us to directly measure channel activity by patching the plasma membrane (Extended Data Fig. 2 and Methods). In brief, MmTPC1 activation requires both membrane

depolarization and the PtdIns(3,5)P₂ ligand (Extended Data Fig. 2b, c). The voltage activation of MmTPC1 is modulated by endolysosomal luminal pH²³, and a lower pH shifts voltage activation towards a more positive potential (Extended Data Fig. 2d, e). In our recordings, MmTPC1 exhibits higher selectivity for Na⁺ than K⁺ and Ca²⁺ (Extended Data Fig. 2f, g); this is different from plant TPC1, which is non-selective^{27,28}.

MmTPC1 structures were determined in the presence and absence of PtdIns(3,5)P₂ to a resolution of 3.2 and 3.4 Å, respectively, using single particle electron cryo-microscopy (cryo-EM) (Fig. 1, Extended Data Figs 3, 4 and Extended Data Table 1). The cryo-EM density maps of both structures are of sufficient quality for *de novo* model building of major parts of the protein (Extended Data Fig. 5). Here we use the higher-resolution PtdIns(3,5)P₂-bound structure for the initial description of the overall structural features. Similar to AtTPC1, each MmTPC1 subunit contains two homologous 6-TM domains (6-TMI and 6-TMII) and two subunits that assemble into a rectangle-shaped functional channel, which is equivalent to a tetrameric *Shaker*-like channel (Fig. 1a, b and Extended Data Fig. 6). Following the same nomenclature as other voltage-gated channels, we labelled the six transmembrane helices within each 6-TM domain as IS1–IS6 and IIS1–IIS6, respectively (Fig. 1c). The transmembrane region of MmTPC1 is domain-swapped; the S1–S4 voltage-sensing domain (VSD) from one 6-TM interacts with the S5–S6 pore domain from the neighbouring 6-TM (Fig. 1b). The pore domain of the second 6-TM contains a luminal loop between IIS5 and pore helix 1 (IIP1) that forms an upright antenna-like β-hairpin; Asn600 and Asn612 on this luminal loop are glycosylated with visible density for the covalently linked *N*-acetylglucosamine moiety of the sugar²⁹ (Fig. 1b, e and Extended Data Fig. 5c).

Multiple cytosolic components within each TPC1 subunit—including the N-terminal H1 helix, the linker between the two 6-TMs and the C-terminal post IIS6 region—assemble into a tightly packed cytosolic domain (Fig. 1d). Despite low sequence homology, the linker between the two 6-TMs adopts the EF-hand domain structure with two EF-hand motifs (EF-1 and EF-2), similar to plant TPC1, and the C-terminal portion of the exceptionally long IS6 serves as the E1 helix (Fig. 1d and Extended Data Fig. 6d). Ca²⁺ is unlikely to bind to the EF motifs of MmTPC1 as these motifs lack essential Ca²⁺-chelating acidic residues (Extended Data Fig. 1). The N-terminal H1 helix is tightly packed with the EF-1 motif and becomes an integral part of the EF-hand domain (Fig. 1d). Compared with plant TPC1 and mammalian TPC2, MmTPC1 has a much longer C-terminal region, which adopts a horseshoe-shaped structure with four α-helices and two β-strands and tightly wraps around the EF-hand domain (Fig. 1d and Extended Data Figs 1, 6d).

The MmTPC1 ion conduction pore, which consists of S5, S6 and two pore helices, adopts a closed conformation in the apo structure and an open conformation in the PtdIns(3,5)P₂-bound structure (Fig. 2a–d). In the apo structure, the four pore-lining S6 helices form

¹Department of Physiology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9040, USA. ²Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas 75390-8816, USA. ³Department of Biophysics, Zhejiang University School of Medicine, Hangzhou 310058, China. ⁴Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9040, USA. ⁵Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9039, USA.

*These authors contributed equally to this work.

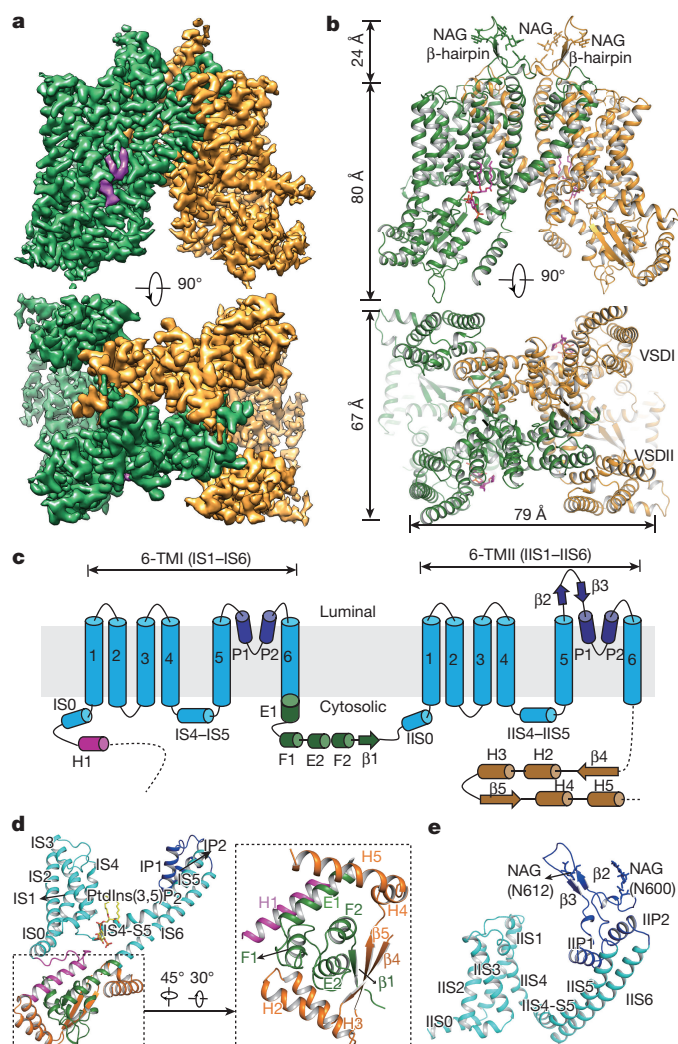


Figure 1 | Overall structure of MmTPC1. **a**, A 3D reconstruction of PtdIns(3,5)P₂-bound (purple density) MmTPC1 dimer with each subunit in individual colour. **b**, Cartoon representation of MmTPC1 in the same orientations as the electron microscopy maps in **a**. *N*-acetylglucosamine (NAG) molecules and PtdIns(3,5)P₂ (purple) are shown as sticks. **c**, Topology and domain arrangement of MmTPC1 subunit. **d**, Structure of the 6-TMI and the soluble domain, with individual elements coloured as in **c**. Inset, zoomed-in view of the cytosolic soluble domain. **e**, Structure of the 6-TMII.

a bundle-crossing at the cytosolic side, with two layers of hydrophobic residues—the Leu317 and Phe321 from each IS6 helix, and the Val684 and Leu688 from each IIS6 helix—that form the constriction points that prevent the passage of hydrated cations (Fig. 2b, c). In the PtdIns(3,5)P₂-bound state, the S6 helices undergo outward movement along with rotational motion (Fig. 2d). Consequently, the constriction-forming residues dilate and rotate away from the central axis, resulting in a much wider opening at the intracellular gate (Fig. 2b–d). In addition, the side chains of four acidic residues (Asp322 from each IS6 and Glu689 from each IIS6) that point tangentially to the pore in the closed structure undergo inward rotation and face the ion conduction pathway in the open state (Fig. 2b–d), generating a ring of negative charges at the gate that could facilitate channel conductance. The molecular mechanism of PtdIns(3,5)P₂-induced channel opening will be discussed later.

The selectivity filter region remains identical in both structures. Two sets of filter residues, Thr280–Ala281–Asn282 (filter I) in 6-TMI and Val647–Asn648–Asn649 (filter II) in 6-TMII, enclose the central ion pathway with different dimensions (Fig. 2e, f). The residues of filter I line the pathway with predominantly main-chain backbone carbonyls

and have atom-to-atom cross distances of about 8 Å (Fig. 2e). The residues of filter II use side chains to generate a much narrower pathway, with two constriction points formed by the Asn648 and Asn649 residues (Fig. 2e). Positioned at the centre of the filter and stabilized by hydrogen-bonding interactions with the filter I backbone carbonyls of Thr280 and Ala281, the Asn648 side chain forms the narrowest point along the filter pathway; it has a cross distance of about 3.7 Å and has the central role in defining the Na⁺ selectivity of MmTPC1 (Fig. 2e, f). Asn648Ala mutation results in a complete loss of Na⁺ selectivity (Fig. 2g and Extended Data Fig. 7). The Asn649 residues are positioned at the luminal entrance of the channel at a wider distance from one another, and Asn649Ala mutation reduces but does not abolish Na⁺ selectivity (Fig. 2g and Extended Data Fig. 7). With an elongated coin-slot-like ion pathway at the filter, Na⁺ ions probably pass through the MmTPC1 filter in a partially hydrated form. The two Asn648 side chains are positioned to provide optimal coordination to stabilize the permeating Na⁺ ion, but are too close to permit the passage of K⁺. Thus, Asn648 forms a simple size sieve to exclude K⁺ or larger ions.

The two VSDs (VSD1 and VSD2) have virtually the same structures as their respective counterparts between the apo and ligand-bound states and, therefore, the higher-resolution PtdIns(3,5)P₂-bound structure will be used in the discussion (Extended Data Fig. 8a, b). Figure 3a provides the numbering of the S4 gating charge residues (R1–R5) from TPCs and other canonical voltage-gated channels for comparison. Although VSD1 contains three arginine residues in IS4 (Arg200, Arg203 and Arg206, at positions R2, R3 and R4, respectively) (Fig. 3b), it lacks some key features of canonical voltage sensors (see Extended Data Fig. 8a legend) and does not contribute to voltage-dependent gating—similar to the VSD1 of plant TPC1²⁵—as shown by the fact that replacing these arginines individually with a neutral residue does not affect the voltage activation of MmTPC1 (Fig. 3c and Extended Data Fig. 8c).

VSD2 contains only two S4 arginines (Arg540 at position R3 and Arg546 at position R5), preserves the key features of a canonical voltage sensor—including the 3₁₀-helix in IIS4 and the conserved gating-charge transfer centre (Fig. 3d)—and is responsible for the voltage gating of MmTPC1. Mutations of Arg540 and Arg546 have a profound but opposite effect on the voltage dependence of the channel. The Arg540Gln mutation stabilizes VSD2 in an activated state and yields a voltage-independent channel that has a linear current–voltage relationship between –100 and 50 mV and can be activated by PtdIns(3,5)P₂ even at hyperpolarization (Fig. 3e). The Arg546Gln mutant, by contrast, can barely be activated by voltage even at high concentrations of PtdIns(3,5)P₂, as if the voltage sensor is trapped in the resting state (Extended Data Fig. 8d).

The MmTPC1 VSD2 adopts an activated conformation with its final voltage-sensing arginine (Arg546) positioned in the gating-charge transfer centre formed by Tyr487 and Glu490 from IIS2 and Asp512 from IIS3, and the other voltage-sensing residues (Arg540 and Gln543) facing the external, luminal side (Fig. 3d). The VSD2 of AtTPC1, with its S4 arginines residing at positions R3–R5, is also responsible for the voltage-gating of the channel, and its structure is in the resting state²⁵. We can, therefore, extrapolate the conformational change of MmTPC1 VSD2 from the activated to the resting state by comparing its structure with that of AtTPC1 (Fig. 3f, g). Except for the S4 helix, the two structures superimpose well. This suggests that upon hyperpolarization the IIS4 of MmTPC1 would slide down by about two helical turns without undergoing structural change in the rest of VSD2, and position its R3 arginine (Arg540) at the gating-charge transfer centre (Fig. 3g). Concurrent with the IIS4 sliding, the IIS4–S5 linker would swing downward and move closer to IIS6. Notably, VSD2 is in the activated state in both the apo and PtdIns(3,5)P₂-bound structures, indicating that the voltage sensor can be activated without opening the channel in MmTPC1.

The bound PtdIns(3,5)P₂ can be unambiguously identified from the electron microscopy density map of the ligand-bound

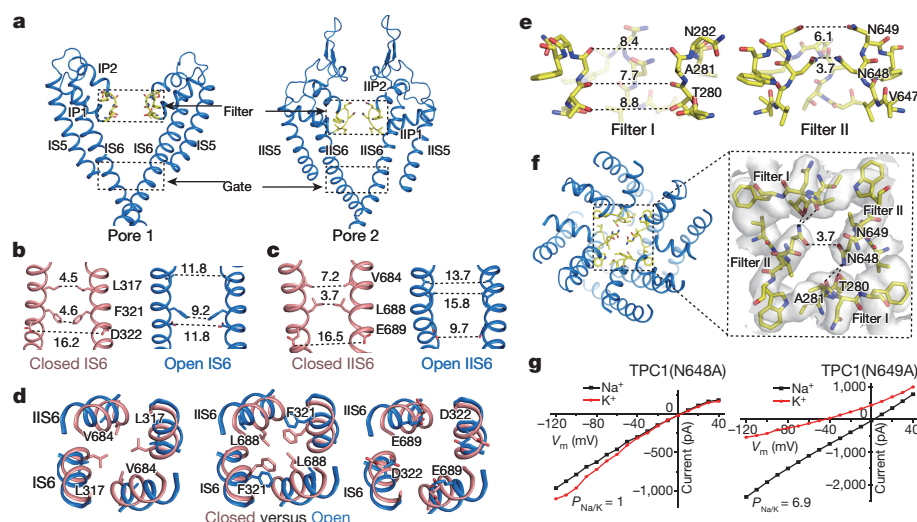


Figure 2 | Ion conduction pore of MmTPC1. **a**, Ion conduction pore, comprising IS5–S6 (pore 1) and IIS5–S6 (pore 2). **b**, **c**, Side view of the bundle-crossing formed by IS6 (**b**) and IIS6 (**c**) in the apo closed (salmon) and PtdIns(3,5)P₂-bound open (blue) states. Numbers are cross distances (in Å) at the constriction points. **d**, Structural comparison of the cytosolic gate between the closed and open states viewed from the cytosolic side in three sections: Leu317/Val684 (left), Phe321/Leu688 (middle) and Asp322/Glu689 (right). **e**, Side view of the selectivity filter formed by

filter I and filter II with the front subunit removed for clarity. Numbers are cross distances (in Å) at the constriction points. **f**, Top view of the selectivity filter. Inset, zoomed-in view of the filter with the stabilization H-bonds for Asn648 (dotted line) and electron microscopy density (grey) shown. **g**, Sample *I*–*V* curves of the filter mutations recorded with high Na⁺ or K⁺ in the bath solution. *P*_{Na/K}, permeability ratio of Na⁺ to K⁺. Original traces are shown in Extended Data Fig. 7. The experiments were repeated five times independently with similar results.

structure (Extended Data Figs 5d, 9a). PtdIns(3,5)P₂ is situated at the junction formed by IS3, IS4 and the IS4–S5 linker of 6-TMI; its inositol 1,3,5-trisphosphate head group is positioned on the cytosolic side and its acyl chains are inserted upright into the membrane (Fig. 4a and Extended Data Fig. 9a). Figure 4b summarizes the protein–ligand interactions, which involve predominantly basic residues from the C terminus of H1, the N terminus of IS3, the IS4–S5 linker and the C-terminal part of IS6. Buried deep in the protein, the two phosphate groups on the C1 and C3 positions of the inositol must be the majority of protein–ligand interactions and probably define the ligand specificity. The C5 phosphate protrudes outwardly away from the ligand-binding pocket, and forms salt bridges with Lys87 and Lys331; the interaction with

Lys331 participates in the coupling between the ligand and IIS6, and has an important role in the ligand activation of the channel. Among all the ligand-interacting residues at the PtdIns(3,5)P₂-binding site, mutations of the residues that predominantly interact with the C3 phosphate—including the three arginines (Arg220, Arg221 and Arg224) on the IS4–S5 linker and Lys331 on IS6—appear to have the most profound effect on PtdIns(3,5)P₂ activation, which illustrates the central role of the C3 phosphate (Extended Data Fig. 9b). In a recent study, the three linker arginines have also been reported to be important for NAADP-mediated Ca²⁺ release³⁰.

To investigate the affinity and specificity of the ligand, we measured the activity of the ligand-dependent channel in excised patches

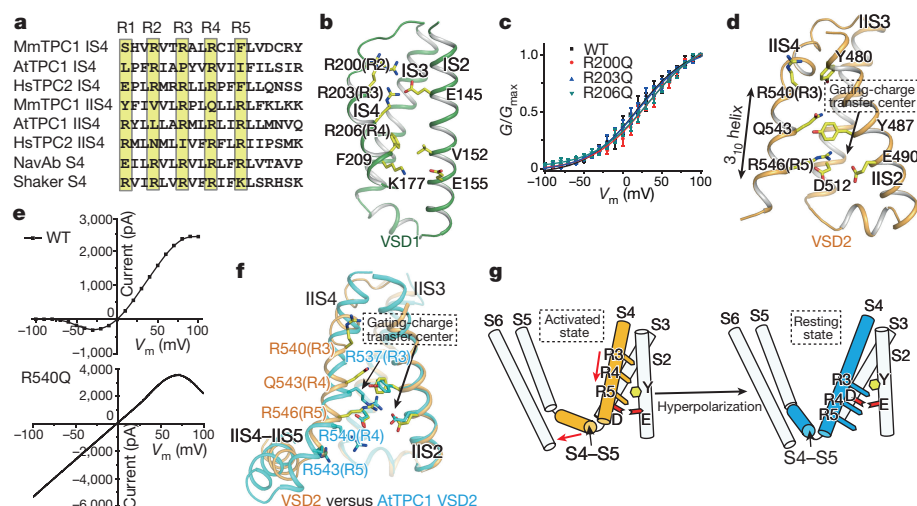


Figure 3 | The voltage-sensing domains. **a**, Partial S4 sequence alignment and arginine registry. NavAb, Nav channel from *Arcobacter butzleri*. **b**, Side view of VSD1 with IS1 omitted for clarity. **c**, *G*/*G*_{max}–*V* curves of wild-type MmTPC1 and IS4 arginine mutations. Sample traces are shown in Extended Data Fig. 8. All data points are mean \pm s.e.m. (*n* = 5 independent experiments). **d**, Side view of VSD2 with IIS1 omitted for clarity. **e**, Sample *I*–*V* curves of wild-type MmTPC1 (obtained from the peak currents at various activation potentials) and Arg540Gln mutant

(obtained by applying a voltage pulse ramp from –100 to 100 mV). Currents were recorded with 2 μ M PtdIns(3,5)P₂ in the pipette and repeated five times independently with similar results. **f**, Structural comparison of VSD2 between the PtdIns(3,5)P₂-bound MmTPC1 (orange) and AtTPC1 (cyan), with S1 helices omitted for clarity. **g**, Cartoon representation of VSD2 conformational change from the activated to resting state. Red arrows indicate the concurrent movements of S4 and S4–S5 linker.

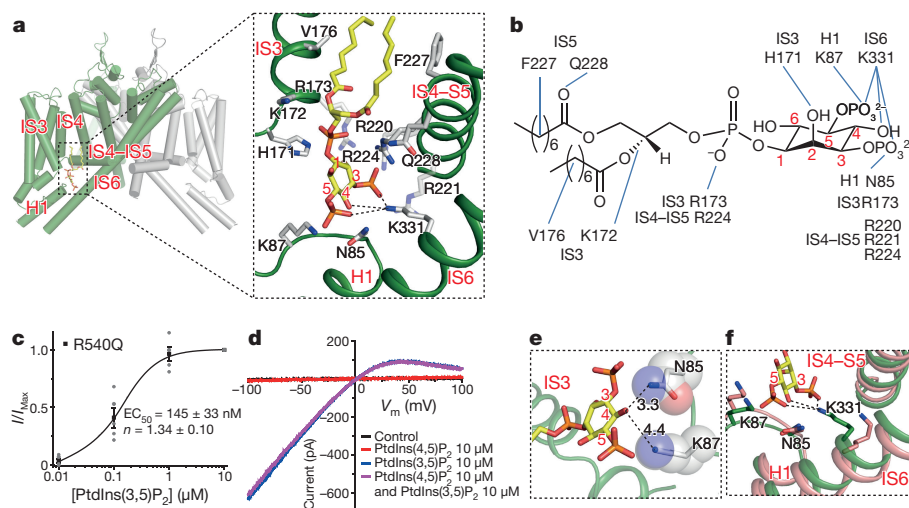


Figure 4 | PtdIns(3,5)P₂ binding in MmTPC1. **a**, PtdIns(3,5)P₂ binding in 6-TM1 of MmTPC1. Inset: zoomed-in view of the PtdIns(3,5)P₂ site. **b**, Schematic of the protein-ligand interactions. **c**, Concentration-dependent PtdIns(3,5)P₂ activation of Arg540Gln mutant at -100 mV. Curve is least square fit to the Hill equation. Data points are mean \pm s.e.m. ($n=5$ independent experiments). Sample *I-V* curves are shown in Extended Data Fig. 9c. **d**, Ligand specificity of MmTPC1 measured

using the Arg540Gln mutant. Sample *I-V* curves were recorded on the same patch with different PtdInsP₂ isoforms. The experiments were repeated five times independently with similar results. **e**, Close proximity between the C4 hydroxyl of PtdIns(3,5)P₂ and the surrounding residues. **f**, Structural comparison at the region around Lys331 between the apo (green) and PtdIns(3,5)P₂-bound MmTPC1 (salmon).

by using the voltage-independent Arg540Gln mutant, which simplifies ligand-dependent gating by eliminating the voltage effect. The mutant also elicits much larger currents, which makes it suitable for inside-out patches. The PtdIns(3,5)P₂-dependent activation of the mutant yielded a half-maximal effective concentration (EC₅₀) of about 145 nM (Fig. 4c and Extended Data Fig. 9c), similar to that of human TPC1 measured in whole lysosome patch²³. The PtdIns(4,5)P₂ isoform cannot activate the channel or inhibit PtdIns(3,5)P₂ activation (Fig. 4d), indicating high lipid specificity of MmTPC1. The lack of PtdIns(4,5)P₂-binding can be explained by the missing C3 phosphate and the close proximity of Asn85 and Lys87 to the C4 hydroxyl group, which sterically excludes the C4 phosphate and thereby prevents the binding of the lipid (Fig. 4e).

Compared to the apo structure, PtdIns(3,5)P₂-binding does not introduce major structural changes around the ligand-binding pocket (Extended Data Fig. 9d), except for one key conformational change on IS6 mediated by Lys331 (Fig. 4f). In the apo state, the Lys331 side chain points away from the ligand-binding pocket. In the presence of PtdIns(3,5)P₂, the Lys331 side chain adopts an extended configuration to form salt bridges with both the C3 and C5 phosphates as well as a hydrogen bond with the C4 hydroxyl, pulling IS6 towards the ligand-binding pocket (Fig. 4f). This movement propagates to the other part of IS6, as well as IIS6, and opens the gate. Lys331 appears to be the only residue that couples IS6 to the bound PtdIns(3,5)P₂ and its mutation to Ala completely abolishes PtdIns(3,5)P₂ activation (Extended Data Fig. 9b).

Our structures demonstrate that PtdIns(3,5)P₂ only binds to the first 6-TM domain and directly introduces conformational changes in IS6 helix, whereas voltage influences only the VSD2 in the second 6-TM domain, the conformational change of which is likely to affect the movement of IIS6 helix (Figs 3, 4). A global structural comparison between the apo and PtdIns(3,5)P₂-bound structures explains the interplay between the two stimuli (Fig. 5). Despite having an activated voltage sensor, the MmTPC1 pore remains closed in the apo structure, implying that PtdIns(3,5)P₂-binding is required to trigger the opening of the gate. Upon PtdIns(3,5)P₂-binding, the ensuing tethering interaction between Lys331 and PtdIns(3,5)P₂ straightens the IS6 helices that are initially bent at the π -helix just below the filter region in the closed state, resulting in the outward dilation and rotation at the bundle crossing (Figs 2d, 5a). The five-residue π -helix is present only in IS6 and may facilitate the helix bending. To open the pore, the IIS6 helices

also have to undergo concurrent outward and rotational movements to accommodate the PtdIns(3,5)P₂-induced conformational change in IS6 helices, particularly the rotation of the two IS6 gating residues with large hydrophobic side chains (Leu317 and Phe321). Consequently, the two IIS6 gating residues (Val684 and L688) also rotate away from the central axis and open the gate (Figs 2d, 5b). The IIS6 motion is hinged around the residue immediately below the filter region, and is propagated to a much larger movement at the C-terminal end of IIS6, which swings upward and makes direct contact with the IIS4-S5 linker. Such motion is permitted only when IIS4 of VSD2 is in the activated, up state. Under hyperpolarized membrane potential, IIS4 is expected to slide downward and push the IIS4-IIS5 linker along with it, occluding the

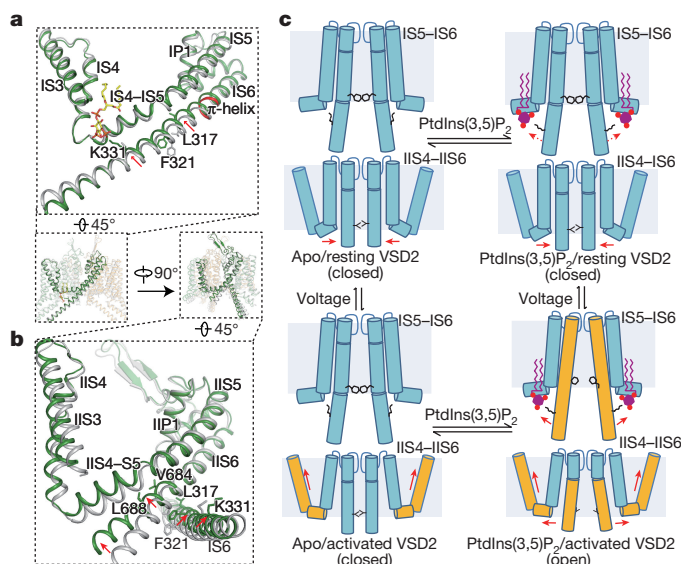


Figure 5 | Gating mechanism of MmTPC1. **a**, **b**, Structural comparison between the apo closed (grey) and PtdIns(3,5)P₂-bound open (green) MmTPC1 with zoomed-in views of the IS3-S6 (**a**) and IIS3-S6 (**b**) regions. Arrows indicate the S6 movements. Key gating residues are shown as sticks. IS6 contains a five-residue π -helix (coloured red). **c**, Working model for voltage-dependent PtdIns(3,5)P₂ activation of MmTPC1. Red arrows mark the direction of the driving force.

space necessary for upward IIS6 movement upon PtdIns(3,5)P₂ activation (Fig. 3f, g). PtdIns(3,5)P₂ can probably still bind MmTPC1 under hyperpolarization, but the resting VSD2 prevents channel opening by blocking the movement of IIS6. Thus, membrane potential modulates the TPC1 channel activity by imposing a voltage-dependent constraint on PtdIns(3,5)P₂ activation and the upward movement of VSD2 under depolarization is a prerequisite for the PtdIns(3,5)P₂-induced gate opening (Fig. 5c).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 November 2017; accepted 19 February 2018.

Published online 21 March 2018.

- Rahman, T. *et al.* Two-pore channels provide insight into the evolution of voltage-gated Ca²⁺ and Na⁺ channels. *Sci. Signal.* **7**, ra109 (2014).
- Yu, F. H. & Catterall, W. A. The VGL-chanome: a protein superfamily specialized for electrical signaling and ionic homeostasis. *Sci. STKE* **2004**, re15 (2004). 10.1126/stke.2532004re15
- Ishibashi, K., Suzuki, M. & Imai, M. Molecular cloning of a novel form (two-repeat) protein related to voltage-gated sodium and calcium channels. *Biochem. Biophys. Res. Commun.* **270**, 370–376 (2000).
- Furuichi, T., Cunningham, K. W. & Muto, S. A putative two pore channel AtTPC1 mediates Ca²⁺ flux in *Arabidopsis* leaf cells. *Plant Cell Physiol.* **42**, 900–905 (2001).
- Patel, S. Function and dysfunction of two-pore channels. *Sci. Signal.* **8**, re7 (2015).
- Grimm, C., Chen, C. C., Wahl-Schott, C. & Biel, M. Two-pore channels: catalyzers of endolysosomal transport and function. *Front. Pharmacol.* **8**, 45 (2017).
- Xu, H. & Ren, D. Lysosomal physiology. *Annu. Rev. Physiol.* **77**, 57–80 (2015).
- Ambrosio, A. L., Boyle, J. A., Aradi, A. E., Christian, K. A. & Di Pietro, S. M. TPC2 controls pigmentation by regulating melanosome pH and size. *Proc. Natl Acad. Sci. USA* **113**, 5622–5627 (2016).
- Bellono, N. W., Escobar, I. E. & Oancea, E. A melanosomal two-pore sodium channel regulates pigmentation. *Sci. Rep.* **6**, 26570 (2016).
- Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **40**, 835–837 (2008).
- Fernández, B. *et al.* Iron overload causes endolysosomal deficits modulated by NAADP-regulated 2-pore channels and RAB7A. *Autophagy* **12**, 1487–1506 (2016).
- Pereira, G. J. *et al.* Nicotinic acid adenine dinucleotide phosphate (NAADP) regulates autophagy in cultured astrocytes. *J. Biol. Chem.* **286**, 27875–27881 (2011).
- Favia, A. *et al.* VEGF-induced neoangiogenesis is mediated by NAADP and two-pore channel-2-dependent Ca²⁺ signaling. *Proc. Natl Acad. Sci. USA* **111**, E4706–E4715 (2014).
- Arndt, L. *et al.* NAADP and the two-pore channel protein 1 participate in the acrosome reaction in mammalian spermatozoa. *Mol. Biol. Cell* **25**, 948–964 (2014).
- Cang, C. *et al.* mTOR regulates lysosomal ATP-sensitive two-pore Na⁺ channels to adapt to metabolic state. *Cell* **152**, 778–790 (2013).
- Grimm, C. *et al.* High susceptibility to fatty liver disease in two-pore channel 2-deficient mice. *Nat. Commun.* **5**, 4699 (2014).
- Sakurai, Y. *et al.* Two-pore channels control Ebola virus host cell entry and are drug targets for disease treatment. *Science* **347**, 995–998 (2015).
- Brailoiu, E. *et al.* Essential requirement for two-pore channel 1 in NAADP-mediated calcium signaling. *J. Cell Biol.* **186**, 201–209 (2009).
- Calcraft, P. J. *et al.* NAADP mobilizes calcium from acidic organelles through two-pore channels. *Nature* **459**, 596–600 (2009).
- Zong, X. *et al.* The two-pore channel TPCN2 mediates NAADP-dependent Ca²⁺-release from lysosomal stores. *Pflügers Arch.* **458**, 891–899 (2009).
- Wang, X. *et al.* TPC proteins are phosphoinositide-activated sodium-selective ion channels in endosomes and lysosomes. *Cell* **151**, 372–383 (2012).
- Jha, A., Ahuja, M., Patel, S., Brailoiu, E. & Muallem, S. Convergent regulation of the lysosomal two-pore channel-2 by Mg²⁺, NAADP, PI(3,5)P₂ and multiple protein kinases. *EMBO J.* **33**, 501–511 (2014).
- Cang, C., Bekele, B. & Ren, D. The voltage-gated sodium channel TPC1 confers endolysosomal excitability. *Nat. Chem. Biol.* **10**, 463–469 (2014).
- Rybalchenko, V. *et al.* Membrane potential regulates nicotinic acid adenine dinucleotide phosphate (NAADP) dependence of the pH- and Ca²⁺-sensitive organellar two-pore channel TPC1. *J. Biol. Chem.* **287**, 20407–20416 (2012).
- Guo, J. *et al.* Structure of the voltage-gated two-pore channel TPC1 from *Arabidopsis thaliana*. *Nature* **531**, 196–201 (2016).
- Kintzer, A. F. & Stroud, R. M. Structure, inhibition and regulation of two-pore channel TPC1 from *Arabidopsis thaliana*. *Nature* **531**, 258–264 (2016).
- Hedrich, R. & Marten, I. TPC1-SV channels gain shape. *Mol. Plant* **4**, 428–441 (2011).
- Guo, J., Zeng, W. & Jiang, Y. Tuning the ion selectivity of two-pore channels. *Proc. Natl Acad. Sci. USA* **114**, 1009–1014 (2017).
- Hooper, R., Churamani, D., Brailoiu, E., Taylor, C. W. & Patel, S. Membrane topology of NAADP-sensitive two-pore channels and their regulation by N-linked glycosylation. *J. Biol. Chem.* **286**, 9141–9149 (2011).
- Patel, S., Churamani, D. & Brailoiu, E. NAADP-evoked Ca²⁺ signals through two-pore channel-1 require arginine residues in the first S4–S5 linker. *Cell Calcium* **68**, 1–4 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Nguyen for manuscript preparation and M. X. Zhu for providing clones of animal TPC genes. Single particle cryo-EM data were collected at the University of Texas Southwestern Medical Center (UTSW) Cryo-Electron Microscopy Facility that is funded by the CPRIT Core Facility Support Award RP170644. We thank D. Nicastro and Z. Chen for facility access and data acquisition. Negatively stained sample screening was performed at UTSW Electron Microscopy core. This work was supported in part by the Howard Hughes Medical Institute (Y.J.) and by grants from the National Institute of Health (GM079179 to Y.J.) and the Welch Foundation (Grant I-1578 to Y.J.). X.B. is supported by the Cancer Prevention and Research Initiative of Texas and Virginia Murchison Linthicum Scholar in Medical Research fund.

Author Contributions J.S., J.G. and Q.C. prepared the samples; J.S., J.G., Q.C. and X.B. performed data acquisition, image processing and structure determination; W.Z. performed electrophysiology; Y.J. supervised the project and revised the manuscript; all authors participated in research design, data analysis and manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Y.J. (youxing.jiang@utsouthwestern.edu) or X.B. (xiaochen.bai@utsouthwestern.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. Mouse TPC1 (MmTPC1, NCBI accession: NM_145853.2) containing a C-terminal thrombin cleavage site followed by a GFP tag and a 10× His tag was cloned into a pEZT-BM vector³¹ and heterologously expressed in HEK293F cells (Life Technologies) using the BacMam system (Thermo Fisher Scientific). The baculovirus was generated in Sf9 cells (Life Technologies) following the standard protocol and used to infect HEK293F cells at a ratio of 1:40 (virus:HEK293F, v/v) and supplemented with 10 mM sodium butyrate to boost protein expression. Cells were cultured in suspension at 37 °C for 48 h and collected by centrifugation at 3,000g. All purification procedures were carried out at 4 °C. The cell pellet was re-suspended in buffer A (20 mM Tris, pH 8.0, 150 mM NaCl) supplemented with a protease inhibitor cocktail (containing 2 µg/ml DNase, 0.5 µg/ml pepstatin, 2 µg/ml leupeptin, and 1 µg/ml aprotinin and 0.1 mM PMSF) and homogenized by sonication on ice. MmTPC1 was extracted with 1% (w/v) *n*-dodecyl- β -D-maltopyranoside (Anatrace) supplemented with 0.2% (w/v) cholesteryl hemisuccinate (Sigma Aldrich) by gentle agitation for 2 h. After extraction, the supernatant was collected after a 60-min centrifugation at 20,000g and incubated with Ni-NTA resin (Qiagen) using gentle agitation. After 2 h, the resin was collected on a disposable gravity column (Bio-Rad). The resin was washed with buffer B (20 mM Tris, pH 8.0, 150 mM NaCl and 0.06% glycol-diosgenin (Anatrace) supplemented with 20 mM imidazole. The washed resin was left on column in buffer B and digested with thrombin (Roche) overnight. After thrombin digestion, the flow-through containing untagged MmTPC1 was collected, concentrated and purified by size exclusion chromatography on a Superdex 200 column (GE Healthcare) pre-equilibrated with buffer B. The peak fraction was pooled and concentrated to 4.7 mg/ml for cryo-EM analysis. To obtain PtdIns(3,5)P₂ bound structure, the protein sample was supplemented with 0.5 mM PtdIns(3,5)P₂ diC8 (Echelon Biosciences) for 30 min on ice before electron microscopy grid preparation.

Electron microscopy data acquisition. The cryo-EM grids were prepared by applying MmTPC1 (4.7 mg/ml, with or without 0.5 mM PtdIns(3,5)P₂) to a glow-discharged Quantifoil R1.2/1.3 300-mesh gold holey carbon grid. Grids were blotted for 4.0 s under 100% humidity at 4 °C before being plunged into liquid ethane using a Mark IV Vitrobot (FEI). Micrographs were acquired on a Titan Krios microscope (FEI) operated at 300 kV with a K2 Summit direct electron detector (Gatan), using a slit width of 20 eV on a GIF-Quantum energy filter. Images were recorded with EPU software (FEI) in super-resolution counting mode with a super resolution pixel size of 0.535 Å. The defocus range was set from −1.5 µm to −3 µm. Each micrograph was dose-fractionated to 30 frames under a dose rate of 4 e[−] per pixel per s, with a total exposure time of 15 s, resulting in a total dose of about 50 e[−] per Å².

Image processing. Micrographs were motion corrected and binned twofold (yielding a pixel size of 1.07 Å per pixel) with MotionCor2³². The CTF parameters of the micrographs were estimated using the GCTF program³³. All other steps of image processing were performed using RELION v2.0^{34,35}. Initially, ~1,000 particles were manually picked from a few micrographs. Class averages representing projections of MmTPC1 in different orientations were selected from the 2D classification of the manually picked particles, and used as templates for automatic particle picking from the full dataset. For the Apo MmTPC1 dataset, 1,411,763 particles were picked from 2,937 micrographs. The particles were extracted and binned 3 times (3.21 Å per pixel). After 2D classification, a total of 1,117,348 particles were finally selected for 3D classification using the AtTPC1 structure as the initial mode. Three of the 3D classes showed good secondary structure features and were selected and re-extracted into the original pixel size of 1.07 Å. After 3D refinement with C2 symmetry imposed, and particle polishing, the resulting 3D reconstructions from ~536,000 particles showed a clear two-fold symmetry with a resolution of 3.5 Å. We then performed a focused 3D classification with density subtraction to improve the density of transmembrane domain³⁶. In this approach, only the density corresponding to the transmembrane domain was kept in each particle image, by subtracting the density for all other parts including the belt-like detergent density from the original particles. The subsequent 3D classification on the modified particles was carried out by applying a mask around the transmembrane domain with all the particle orientations fixed at the value determined in the initial 3D refinement. After this round of classification, one class (~43,000 particles) showed better density in the transmembrane domain. The corresponding particles before density subtraction from this class were selected and 3D-refined, yielding an electron microscopy map of 3.4 Å for the entire channel.

The data for MmTPC1 in the presence of PtdIns(3,5)P₂ were processed similarly to that of apo MmTPC1. In brief, 941,754 particles were picked from a total of

2,348 micrographs. After 2D classification, 620,307 particles were selected for 3D classification. Three classes with a total of ~245,000 particles were selected and combined for 3D auto-refinement, which resulted in a map with an overall resolution of 3.3 Å. One round of 3D classification was then performed by focusing on the transmembrane domain. One class (~83,000 particles) showed better density in the transmembrane domain and was selected for final 3D refinement, yielding an electron microscopy map of 3.2 Å. All resolutions were estimated by applying a soft mask around the protein density and the gold-standard Fourier shell correlation (FSC) = 0.143 criterion. ResMap was used to calculate the local resolution map³⁷.

Model building, refinement and validation. *De novo* atomic model buildings were conducted in Coot³⁸. Amino acid assignment was achieved based mainly on the clearly defined densities for bulky residues (Phe, Trp, Tyr and Arg). Real-space model refinement was performed in Phenix³⁹. Models were validated using previously described methods, to avoid overfitting^{40,41}. The final structure models for both apo and PtdIns(3,5)P₂-bound states include residues 66–701 and residues 709–795. Residues 1–65, 702–708 and 796–817 are disordered and not modelled. The statistics of the geometries of the models were generated using MolProbity⁴². All the figures were prepared in PyMol⁴³ or Chimera⁴⁴. Programs used for model building, refinement and validation are compiled by SGrid⁴⁵.

Electrophysiology. In human TPC2, the Leu11Ala and Leu12Ala mutations at the N-terminal targeting sequence have previously been shown to promote channel expression and trafficking to the plasma membrane of the HEK293 cell, enabling channel activity measurement using patch clamp^{22,46}. We therefore also introduced the equivalent mutations (Leu11Ala and Ile12Ala) to MmTPC1. HEK293 cells overexpressed with the Leu11Ala/Ile12Ala mutant of MmTPC1 elicited much larger whole-cell currents than those expressed with wild-type MmTPC1 (Extended Data Fig. 2a). Therefore, the Leu11Ala/Ile12Ala mutant was used and considered as the wild-type channel in all our recordings. All other mutations in our experiments were generated on the background of this plasma-membrane-targeting MmTPC1. With the channels targeted to the plasma membrane, the extracellular side is equivalent to the luminal side of TPC1 in endosomes or lysosomes. MmTPC1 and its mutants were cloned into pCGFP-EU vector⁴⁷. About 2 µg of the plasmid containing the C-terminal GFP-tagged MmTPC1 or its mutant was transfected into HEK293 cells grown in a six-well tissue culture dish using Lipofectamine 2000 (Life Technology). Forty-eight hours after transfection, cells were dissociated by trypsin treatment and kept in complete serum-containing medium and re-plated on 35-mm tissue culture dishes in a tissue culture incubator until recording.

Patch clamp in whole-cell configuration was used to measure channel activity in most of the experiments except the measurements of ligand affinity and specificity, which were recorded in excised patches (inside-out patches) using the voltage-independent Arg540Gln mutant. This mutant channel can be activated solely by PtdIns(3,5)P₂ and also yields much larger plasma membrane currents, which makes it more amenable for inside-out patches. The standard intracellular solution contained (in mM): 145 sodium methanesulfonate (Na-MS), 5 NaCl, 4 MgCl₂, 1 EGTA, 10 HEPES buffered with Tris, pH = 7.4. The extracellular solution contained (in mM): 145 Na-MS, 5 NaCl, 1 MgCl₂, 1 CaCl₂, 10 HEPES buffered with Tris, pH = 7.4. Various concentrations of PtdIns(3,5)P₂ as specified in each experiment were added to the intracellular solutions to activate the channel. For patches in whole-cell configuration, the intracellular solution was in the pipette and the extracellular solution was in the bath; the solution arrangement was reversed for the inside-out patches. The lipid ligands used in our studies are phosphatidylinositol-3,5-bisphosphate diC8 (PtdIns(3,5)P₂ diC8, Echelon) and phosphatidylinositol-4,5-bisphosphate diC8 (PtdIns(4,5)P₂ diC8, Echelon).

The data were acquired using an AxoPatch 200B amplifier (Molecular Devices) and a low-pass analogue filter set to 1 kHz. The current signal was sampled at a rate of 20 kHz using a Digidata 1322A digitizer (Molecular Devices) and further analysed with pClamp 9 software (Molecular Devices). Patch pipettes were pulled from borosilicate glass (Harvard Apparatus) and heat-polished to a resistance of 3–5 MΩ. After the patch pipette attached to the cell membrane, a giga-seal (>10 GΩ) was formed by gentle suction. The whole-cell configuration was formed by short zap or suction to rupture the patch. The inside-out configuration was formed by pulling the pipette away from the cell, and the pipette tip was exposed to the air for a short period in some cases. The holding potential was set to −70 mV. To generate *G*/*G*_{max} versus *V* curves (*G* = *I*/*V*), the membrane was stepped from the holding potential (−70 mV) to various testing potentials (−100 mV to 100 mV) for 1 s and then stepped to −70 mV (Extended Data Fig. 2b). The peak tail currents were used to plot the *G*–*V* curve. *G*_{max} was obtained from the peak tail current at 100 mV testing potential. *V*_{1/2} and *Z* values were obtained from the fits of data with Boltzmann equation, in which *V*_{1/2} is the voltage at which the channels have reached half of their maximum fraction open and *Z* is the apparent valence of voltage dependence. The same protocol was used to obtain current and voltage

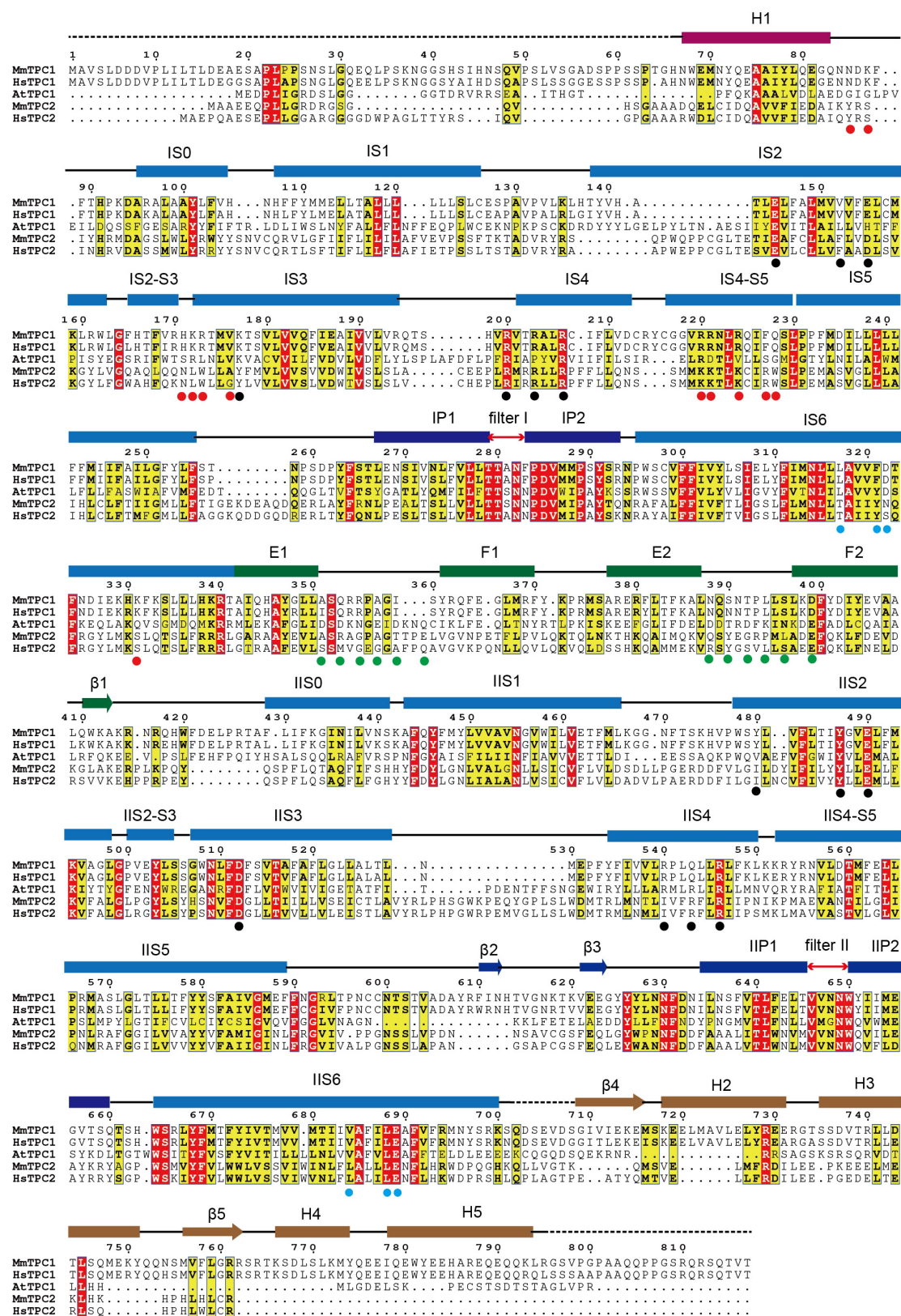
relationships (I – V curve) of the wild-type MmTPC1 (Fig. 3e, top trace), except that the peak current at each testing potential was used to generate the I – V curve. For voltage-independent Arg540Gln mutant, the holding potential was set to 0 mV, and the current and voltage relationship (I – V curve, Fig. 3e, bottom trace) was obtained directly by using voltage pulses ramp from –100 to 100 mV over 800-ms duration.

For measuring ion selectivity of MmTPC1 and its mutants in whole-cell patches, 10 μ M PtdIns(3,5)P₂ was included in intracellular (pipette) solution to fully activate the channel. The membrane potential was stepped from the holding potential (–70 mV) to 100 mV for 1 s to activate the channels, and then stepped to various testing potentials (–120 mV to 4 mV) for 1 s (Extended Data Fig. 2f). The peak tail currents at various testing potentials were plotted to determine the reversal potential (V_{rev}). To measure the relative permeability between Na⁺ and K⁺, the extracellular (bath) solution (in mM) was changed to 145 K-MS, 5 NaCl, 1 MgCl₂, 1 CaCl₂, 10 HEPES buffered with Tris, pH 7.4. To measure the relative permeability between Na⁺ and Ca²⁺, the extracellular solution (in mM) was changed to 95 Ca-(MS)₂, 5 CaCl₂, 10 HEPES buffered with Tris, pH 7.4. The ion permeability ratios were calculated with the equations: $P_{\text{Na}}/P_{\text{K}} = [K]_{\text{o}}/([Na]_{\text{i}}\exp(V_{\text{rev}}F/RT) - [Na]_{\text{o}})$ and $P_{\text{Na}}/P_{\text{Ca}} = 4[Ca]_{\text{o}}/([Na]_{\text{i}}\exp(V_{\text{rev}}F/RT)(1 + \exp(V_{\text{rev}}F/RT)))$, in which V_{rev} is the reverse potential, F is Faraday's constant, R is the gas constant, T is the absolute temperature, o is extracellular and i is intracellular.

All electrophysiological recording were repeated at least five times using different patches. Most data points shown are mean \pm s.e.m. ($n = 5$ independent experiments).

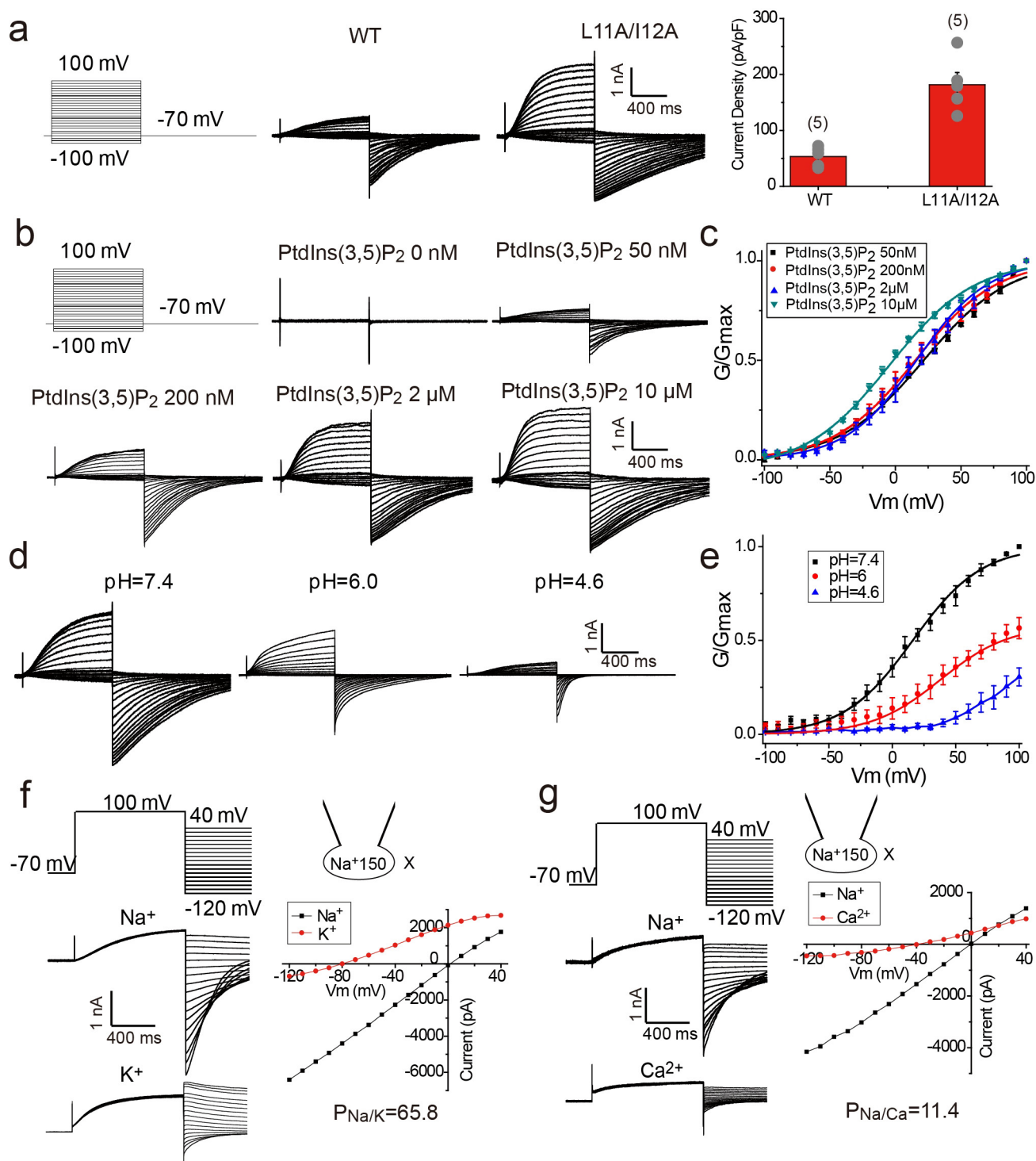
Data availability. The cryo-EM density maps of the MmTPC1 have been deposited in the Electron Microscopy Data Bank under accession number EMD-7434 for the apo state, and accession number EMD-7435 for the PtdIns(3,5)P₂-bound state. Atomic coordinates have been deposited in the RCSB Protein Data Bank under accession number 6C96 for the apo state, and accession number 6C9A for the PtdIns(3,5)P₂-bound state. Source Data for Fig. 3c and Extended Data Fig. 2c, e are available in the online version of the paper.

31. Morales-Perez, C. L., Noviello, C. M. & Hibbs, R. E. Manipulation of subunit stoichiometry in heteromeric membrane proteins. *Structure* **24**, 797–805 (2016).
32. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
33. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
34. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
35. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
36. Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ -secretase. *eLife* **4**, e11182 (2015).
37. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
38. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
39. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
40. Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
41. DiMaio, F., Zhang, J., Chiu, W. & Baker, D. Cryo-EM model validation using independent map reconstructions. *Protein Sci.* **22**, 865–868 (2013).
42. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
43. Schrödinger, L. The PyMOL Molecular Graphics System, Version 1.8. (Schrödinger, 2015).
44. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
45. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
46. Brailoiu, E. *et al.* An NAADP-gated two-pore channel targeted to the plasma membrane uncouples triggering from amplifying Ca²⁺ signals. *J. Biol. Chem.* **285**, 38511–38516 (2010).
47. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14**, 673–681 (2006).



Extended Data Figure 1 | Sequence alignment of MmTPC1, HsTPC1, AtTPC1, MmTPC2 and HsTPC2. Secondary structure assignments are based on the structure of PtdIns(3,5) P_2 -bound MmTPC1. Red dots mark the ligand-binding residues; black dots mark the S4 arginine residues

and residues at the gating-charge transfer centre; cyan dots mark the key S6 gating residues; green dots mark the residues predicted to participate in Ca^{2+} coordination in EF-hand domains of AtTPC1. MmTPC1 and AtTPC1 share about 25% sequence identity.

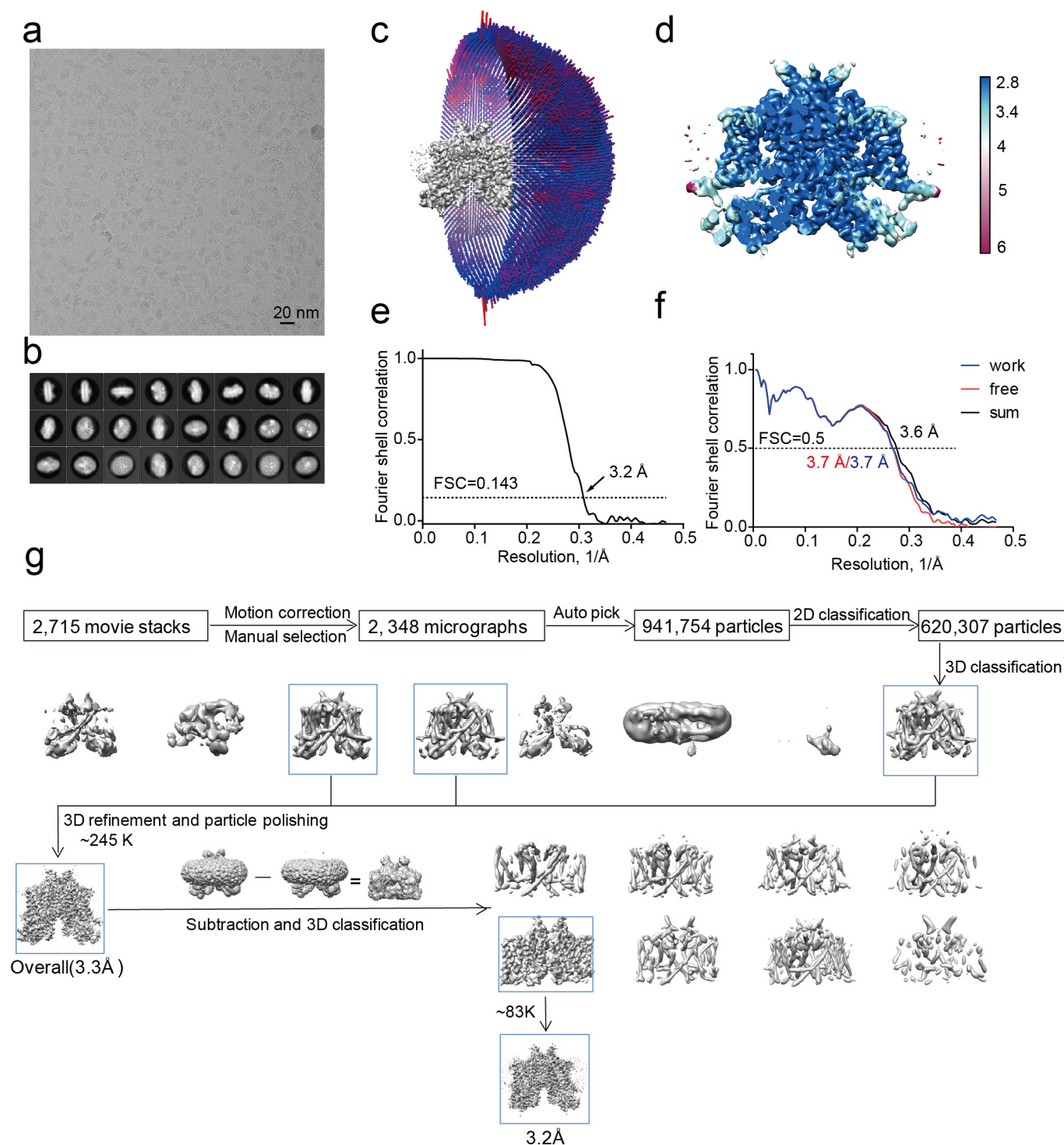


Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Gating and selectivity of MmTPC1.

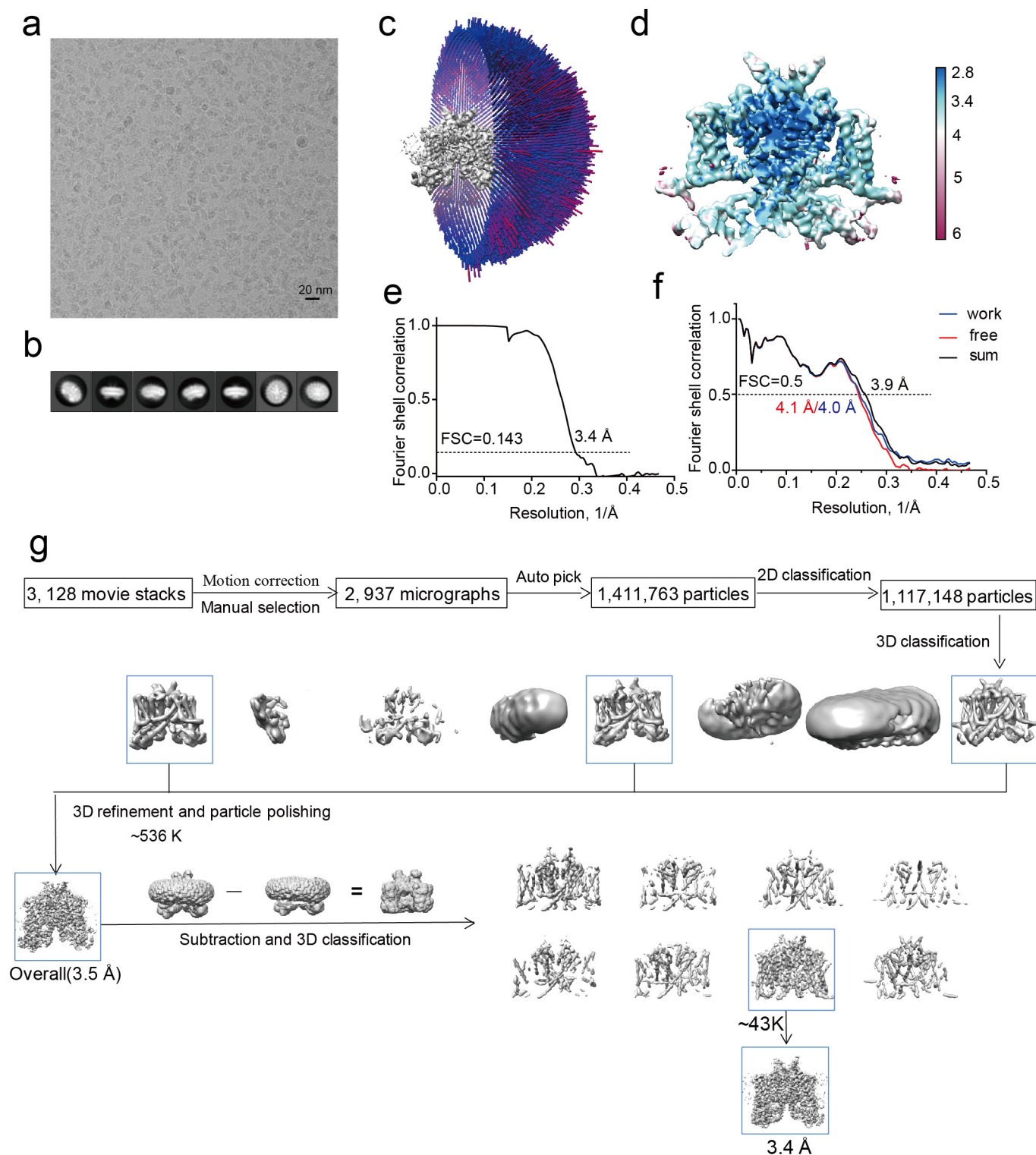
a, Sample traces and current density (current/capacitance) of wild-type MmTPC1 and the L11A/I12A mutant of MmTPC1, recorded in the whole-cell configuration with 100 μ M PtdIns(3,5)P₂ in the pipette (cytosolic). The experiments were repeated five times independently with similar results. Data points for current density are mean \pm s.e.m. ($n=5$ independent experiments). The L11A/I12A mutant elicited much larger whole-cell currents and was therefore used as the wild-type channel in all recordings. The extracellular side of MmTPC1 in plasma membrane is equivalent to the luminal side of MmTPC1 in lysosomes. **b**, Sample traces of PtdIns(3,5)P₂-dependent voltage activation of MmTPC1. Whole-cell currents were recorded with varying PtdIns(3,5)P₂ concentrations in the pipette (cytosolic) at pH 7.4. The experiments were repeated five times independently with similar results. **c**, G/G_{\max} - V curves of MmTPC1 at various PtdIns(3,5)P₂ concentrations. Boltzmann fit yields $V_{1/2}$ (mV) = 21.6 ± 1.2 , 15.2 ± 1.0 , 16.1 ± 0.9 and -2.0 ± 1.0 , and $Z = 0.78 \pm 0.04$, 0.82 ± 0.03 , 0.89 ± 0.02 and 0.84 ± 0.05 for voltage activation in 0.05, 0.2, 2.0 and 10 μ M cytosolic PtdIns(3,5)P₂, respectively, in which $V_{1/2}$ is the membrane potential for half maximum activation and

Z is apparent valence. All data points are mean \pm s.e.m. ($n=5$ independent experiments). **d**, Luminal pH modulates the voltage activation of MmTPC1. Whole-cell currents of MmTPC1 recorded in the presence of 2 μ M cytosolic PtdIns(3,5)P₂ with a varying luminal (bath) pH of 7.4, 6.0 or 4.6. Sample traces were obtained from the same patch. The experiments were repeated five times independently with similar results. **e**, G/G_{\max} - V curves of MmTPC1 at various luminal pH values. Boltzmann fit yields $V_{1/2} = 16.2 \pm 0.8$ mV, $Z = 0.91 \pm 0.02$ at pH 7.4, $V_{1/2} = 38.2 \pm 1.2$ mV, $Z = 0.95 \pm 0.02$ at pH 6.0. All data points were normalized against G_{\max} obtained at 100 mV activation voltage and pH 7.4. All data points are mean \pm s.e.m. ($n=5$ independent experiments). **f**, Sample traces of whole-cell currents with 150 mM Na⁺ in the pipette solution, and either 150 mM Na⁺ or 145 mM K⁺ and 5 mM Na⁺ in the bath solution, and the I - V curves generated from the tail currents of the sample traces. **g**, Sample traces of whole-cell currents with 150 mM Na⁺ in the pipette solution and 150 mM Na⁺ or 100 mM Ca²⁺ in the bath solution, and the I - V curves generated from the tail currents of the sample traces. Data in **f** and **g** were recorded with 10 μ M PtdIns(3,5)P₂ in the pipette at pH 7.4 and both experiments were repeated five times independently with similar results.



Extended Data Figure 3 | Structure determination of PtdIns(3,5)P₂-bound MmTPC1. **a**, Representative electron micrograph of PtdIns(3,5)P₂-bound MmTPC1; 2,348 micrographs were used for structure determination. **b**, 2D class averages. **c**, Euler angle distribution of particles used in the final 3D reconstruction, with the heights of the cylinders corresponding to the number of particles. **d**, Final density maps

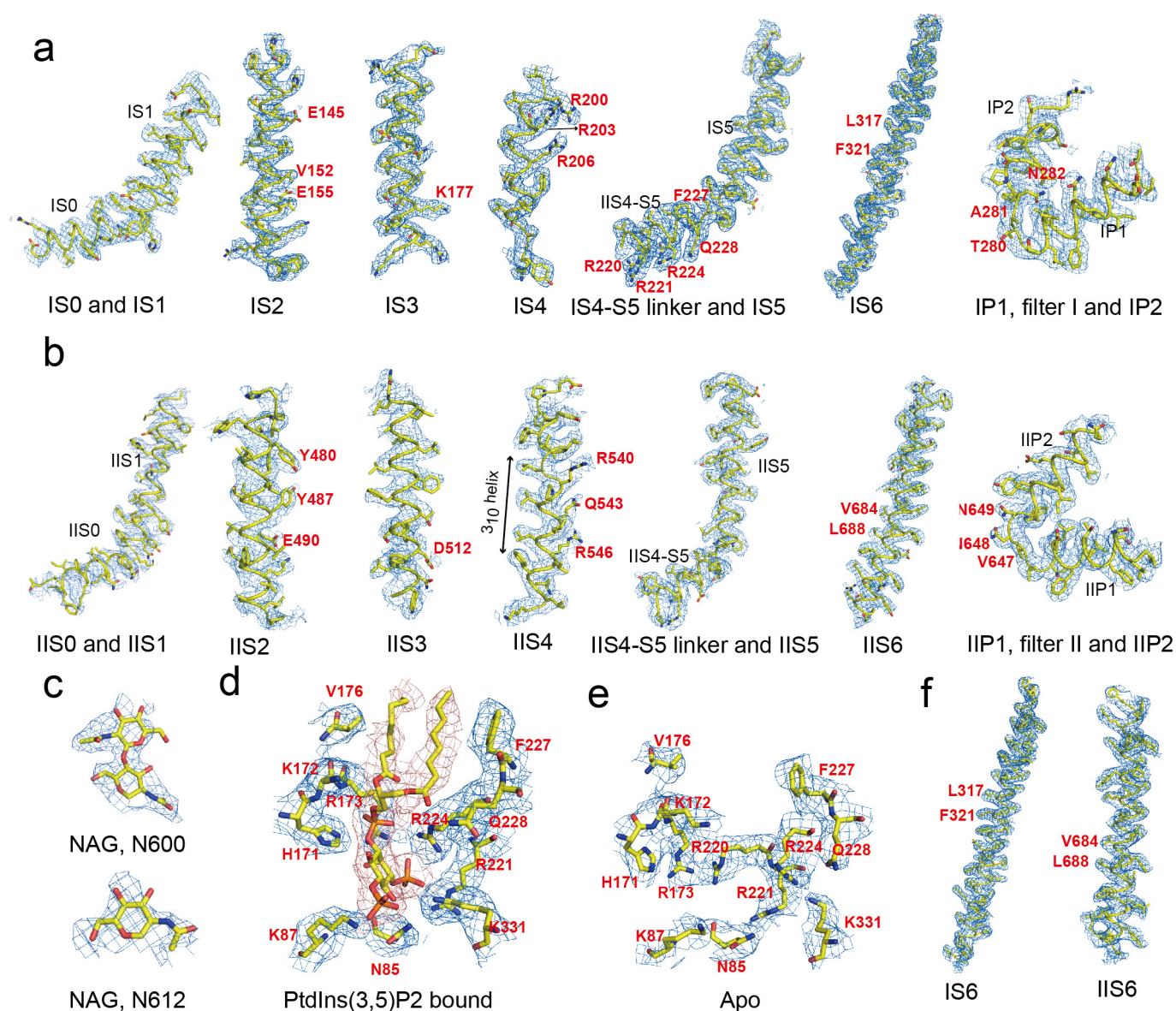
coloured by local resolution. **e**, Gold-standard FSC curves of the final 3D reconstructions. **f**, FSC curves for cross-validation between the models and the maps. Curves for model versus summed map in black (sum), for model versus half map in blue (work) and for model versus half map not used for refinement in red (free). **g**, Flowchart of electron microscopy data processing for PtdIns(3,5)P₂-bound MmTPC1 particles.



Extended Data Figure 4 | Structure determination of apo MmTPC1.

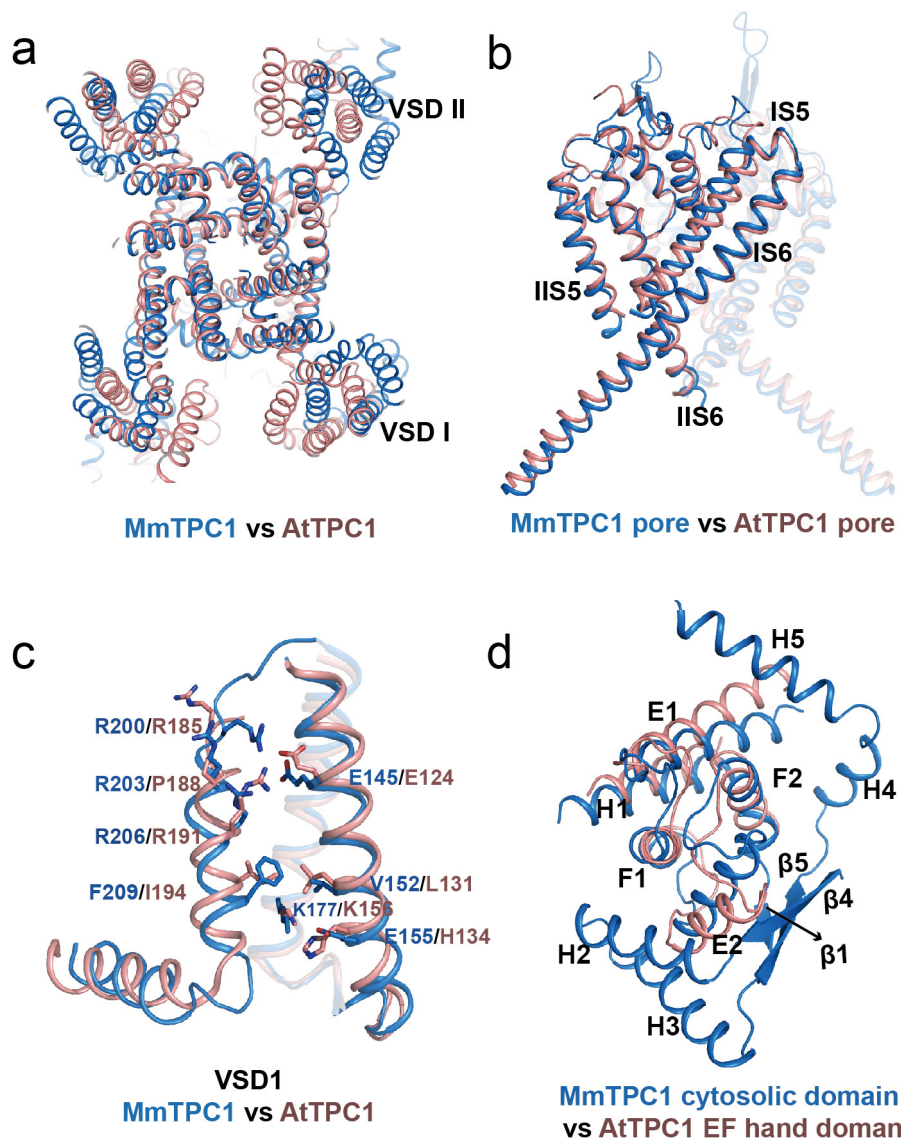
a, Representative electron micrograph of apo MmTPC1; 2,937 micrographs were used for structure determination. **b**, 2D class averages. **c**, Euler angle distribution of particles used in the final 3D reconstruction, with the heights of the cylinders corresponding to the number of particles. **d**, Final density maps coloured by local resolution. **e**, Gold-standard FSC

curves of the final 3D reconstructions. **f**, FSC curves for cross-validation between the models and the maps. Curves for model versus summed map in black (sum), for model versus half map in blue (work) and for model versus half map not used for refinement in red (free). **g**, Flowchart of electron microscopy data processing for apo MmTPC1 particles.



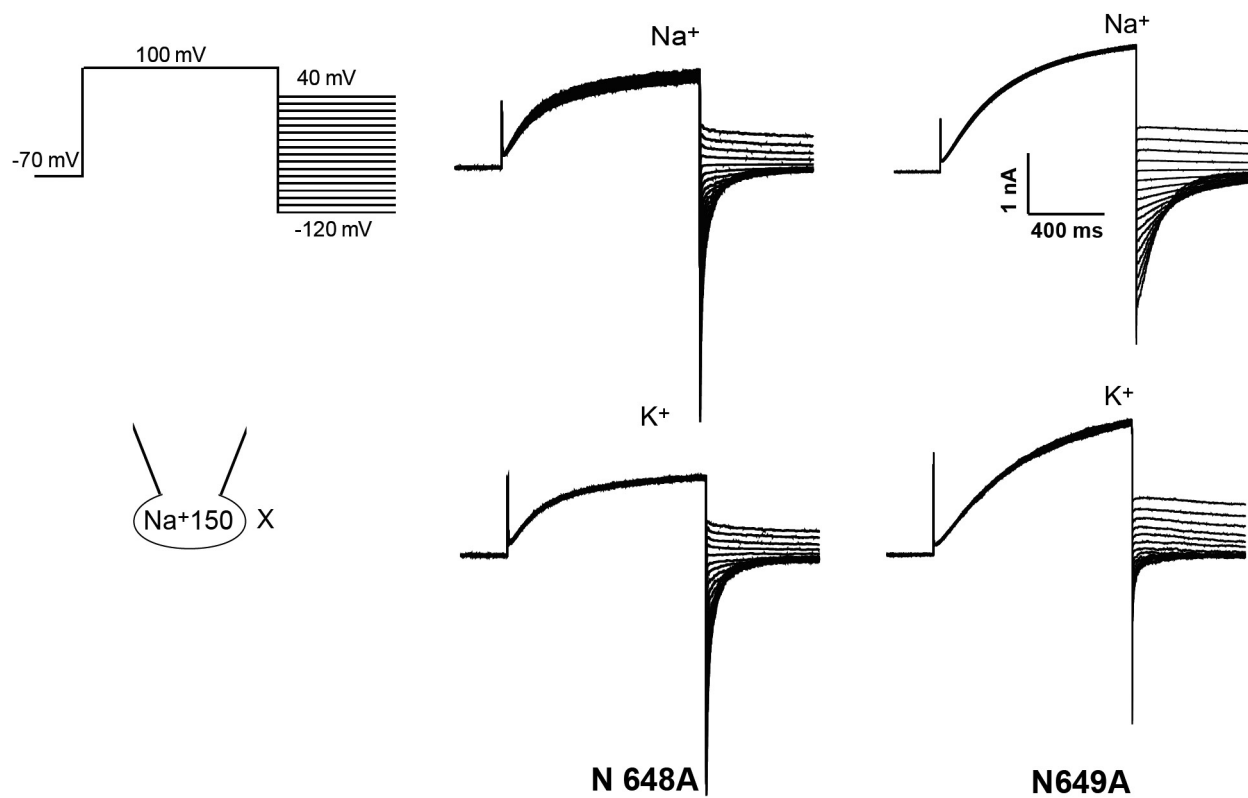
Extended Data Figure 5 | Sample electron microscopy density maps (blue mesh) for MmTPC1. a–d, Sample electron microscopy density maps for various parts of PtdIns(3,5)P₂-bound MmTPC1: IS1–IS6 and filter I (a), IIS1–IIS6 and filter II (b), NAGs of Asn600 and Asn612 (c), and PtdIns(3,5)P₂-binding site (d). The maps are low-pass filtered to 3.2 Å

and sharpened with a temperature factor of -105 \AA^2 . e, f, Sample electron microscopy density maps for the key parts of apo MmTPC1: ligand binding site (e) and S6 helices (f). The maps are low-pass filtered to 3.4 Å and sharpened with a temperature factor of -98.5 \AA^2 . Residues discussed in main text are labelled in red.



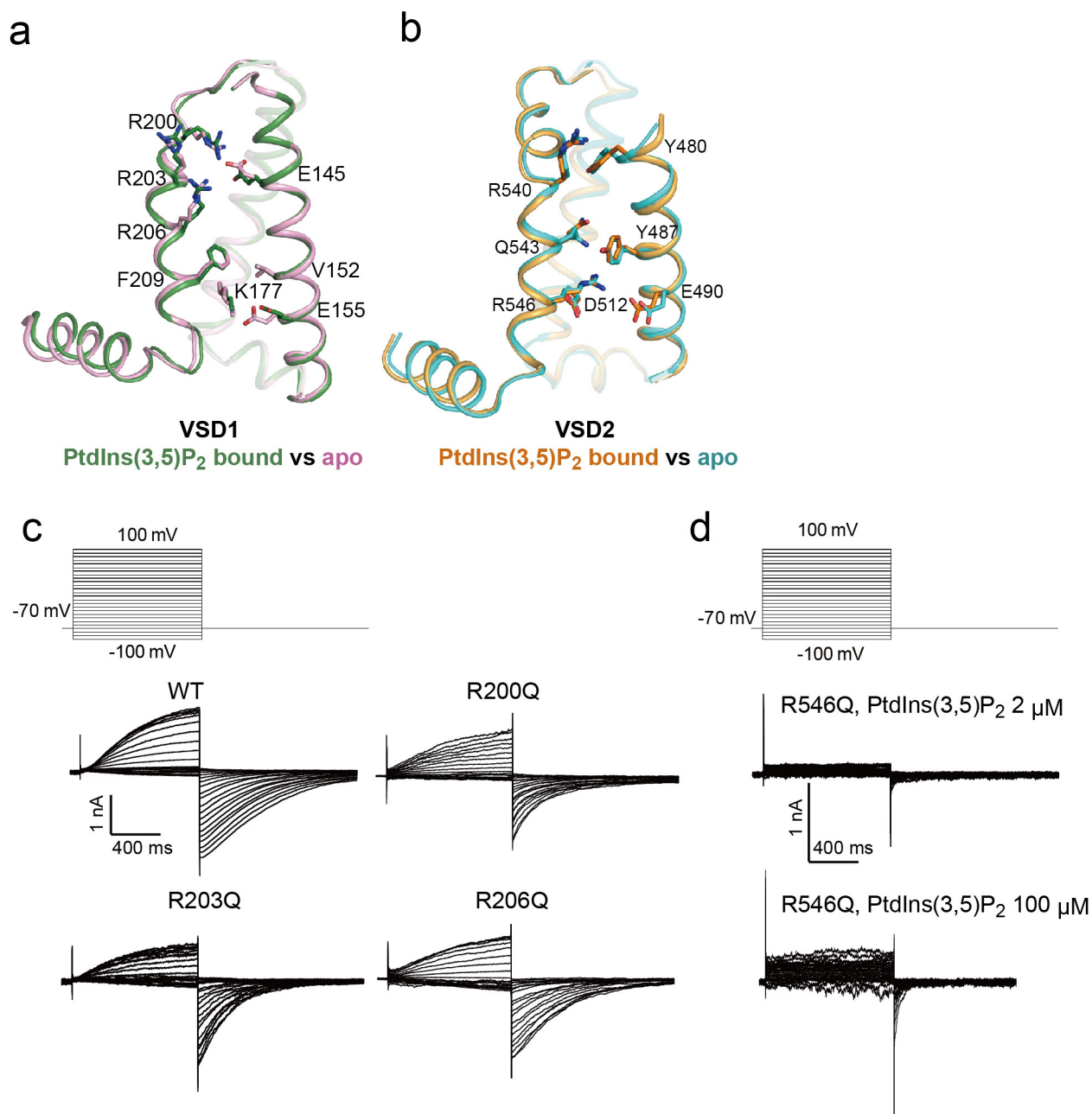
Extended Data Figure 6 | Structure comparison between MmTPC1 and AtTPC1. **a**, Superposition of the overall structures of MmTPC1 (blue) and AtTPC1 (salmon). **b**, Superposition of the pore regions. **c**, Superposition of

VSD1 domains. The comparison of the VSD2 domains is shown in Fig. 3f. **d**, Superposition of cytosolic soluble domains.



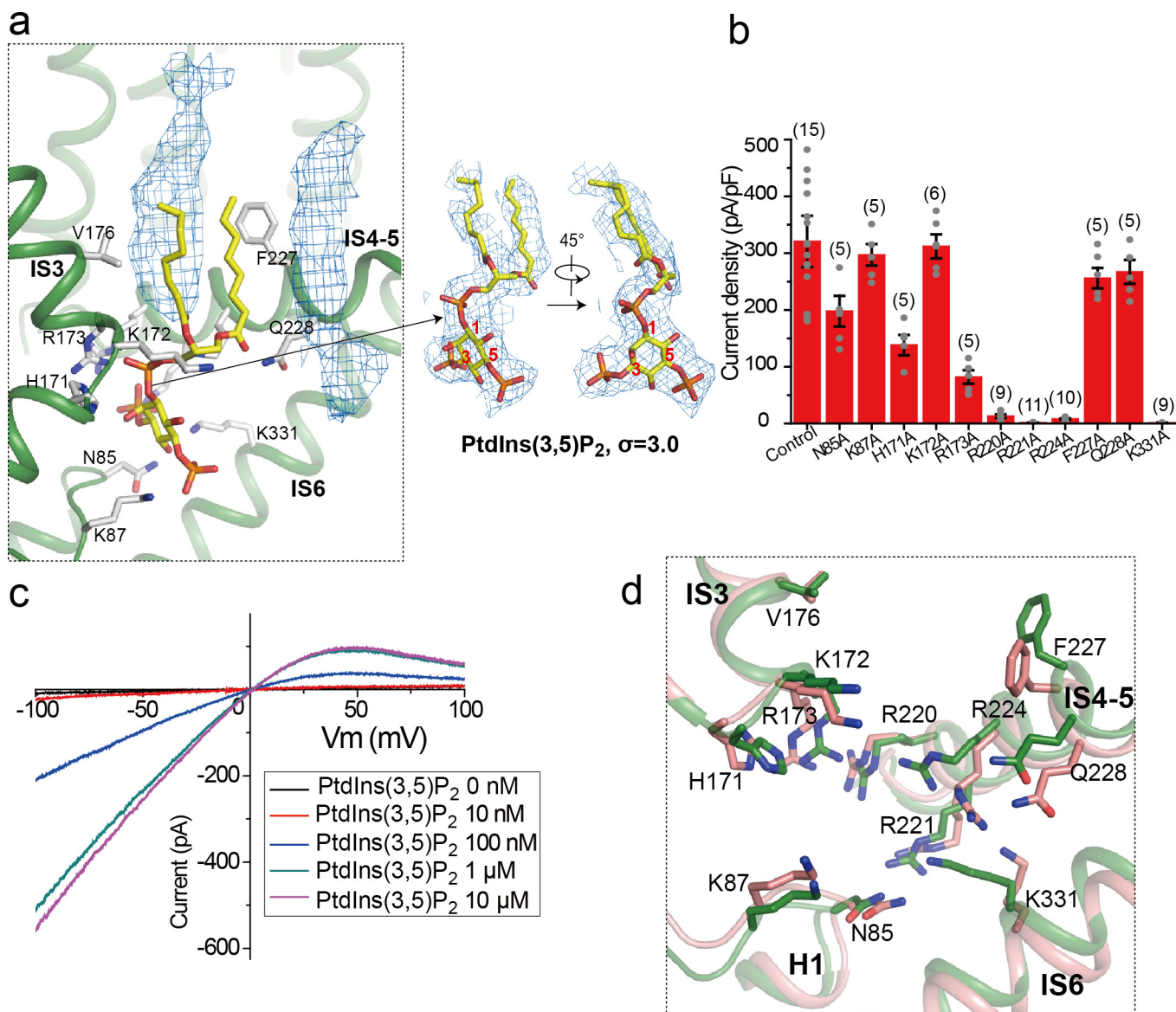
Extended Data Figure 7 | Sample traces of whole-cell currents for Asn648Ala and Asn649Ala filter mutants. The pipette solution contained 150 mM Na⁺ and the bath solution contained 150 mM Na⁺, or 145 mM

K⁺ and 5 mM Na⁺. The tail currents were used to generate the *I*-*V* curves shown in Fig. 2g. The experiments were repeated five times independently with similar results.



Extended Data Figure 8 | Voltage-sensing domains. **a**, Superimposition of MmTPC1 VSD1 structures in the PtdIns(3,5)P₂-bound (green) and apo (pink) states with S1 helices removed for clarity. The MmTPC1 VSD1 lacks some key features of canonical voltage sensors: the conserved aromatic residue on S2 and acidic residue on S3 that form the gating-charge transfer centre become Val152 and Lys177, respectively, in MmTPC1; the conserved basic residue at the R5 position becomes Phe209 in MmTPC1; no arginine from IS4 is positioned in the gating-charge transfer centre. **b**, Superimposition of MmTPC1 VSD2 structures in the

PtdIns(3,5)P₂-bound (orange) and apo (cyan) states. **c**, Sample traces of voltage activation of MmTPC1 and its IS4 arginine mutations, recorded in whole-cell configuration with 2 μM PtdIns(3,5)P₂ in the pipette. Peak tail currents were used to generate the G/G_{\max} - V curves shown in Fig. 3c. The experiments were repeated five times independently with similar results. **d**, Sample traces of voltage activation of Arg546Gln mutation, recorded in whole-cell configuration with 2 μM and 100 μM PtdIns(3,5)P₂ in the pipette. The experiments were repeated five times independently with similar results.



Extended Data Figure 9 | PtdIns(3,5)P₂-binding in MmTPC1. **a**, Model of bound PtdIns(3,5)P₂ (left) and its electron microscopy density (right). Density of two other membrane lipid molecules (blue mesh in left panel) was also observed near PtdIns(3,5)P₂ in the structure. **b**, Current density of mutations at the PtdIns(3,5)P₂-binding site measured at −100 mV in whole-cell recordings. All mutants were generated on the background of Arg540Gln mutant, which was used as control. All data points are mean ± s.e.m. with the number of independent experiments for each

mutant shown in parentheses. **c**, Sample *I*–*V* curves of Arg540Gln mutant recorded in excised patches with varying concentrations of PtdIns(3,5)P₂ in the bath (cytosolic). The experiments were repeated five times independently with similar results. Currents at −100 mV were used to generate the concentration-dependent PtdIns(3,5)P₂ activation curve shown in Fig. 4c. *I*_{max} is the current recorded at −100 mV with 10 μM PtdIns(3,5)P₂ in the bath. **d**, Structural comparison at the ligand-binding site between the PtdIns(3,5)P₂-bound (green) and apo (salmon) states.

Extended Data Table 1 | Cryo-EM data collection and model statistics

	PtdIns(3,5)P ₂ - bound MmTPC1 (EMD-7435) (PDB 6C9A)	Apo MmTPC1 (EMD-7434) (PDB 6C96)
Data collection and processing		
Magnification	46730	46730
Voltage (kV)	300	300
Electron exposure (e ⁻ /Å ²)	~50	~50
Defocus range (μm)	-1.5 to -3.0	-1.5 to -3.0
Pixel size (Å)	1.07	1.07
Symmetry imposed	C2	C2
Initial particle images (no.)	941,754	1,260,054
Final particle images (no.)	82,819	42,870
Map resolution (Å)	3.2	3.4
FSC threshold	0.143	0.143
Refinement		
Initial model used (PDB code)	5E1J	5E1J
Model resolution (Å)	3.2	3.4
FSC threshold	0.143	0.143
Map sharpening <i>B</i> factor (Å ²)	-105.07	-98.52
Model composition		
Non-hydrogen atoms	12182	12090
Protein residues	12004	12004
Ligands	178	86
<i>B</i> factors (Å ²)		
Protein	80.93	98.57
Ligand	67.59	81.03
R.m.s. deviations		
Bond lengths (Å)	0.009	0.010
Bond angles (°)	1.286	1.309
Validation		
MolProbity score	1.41	1.5
Clashscore	2.79	3.5
Poor rotamers (%)	0	0.3
Ramachandran plot		
Favored (%)	94.99%	94.85%
Allowed (%)	5.01%	5.15%
Disallowed (%)	0	0

CORRIGENDUM

doi:10.1038/nature25997

Corrigendum: Commensal bacteria make GPCR ligands that mimic human signalling molecules

Louis J. Cohen, Daria Esterhazy, Seong-Hwan Kim, Christophe Lemetre, Rhiannon R. Aguilar, Emma A. Gordon, Amanda J. Pickard, Justin R. Cross, Ana B. Emiliano, Sun M. Han, John Chu, Xavier Vila-Farres, Jeremy Kaplitt, Aneta Rogoz, Paula Y. Calle, Craig Hunter, J. Kipchirchir Bitok & Sean F. Brady

Nature **549**, 48–53 (2017); doi:10.1038/nature23874

In this Article, the description in the Methods of the human stool samples used for the analysis presented in Extended Data Fig. 9 was incomplete. All samples were collected with informed consent under protocol numbers 09-067, 09-141 and 06-107 at Memorial Sloan Kettering Cancer Center (MSKCC) and stored as part of a biospecimen repository. All sample processing was done at MSKCC. All patients underwent a bone marrow transplant which was standard of care and part of an observational study (not a clinical trial, as originally stated in the Methods). Patient age, gender and transplant indication are provided in the table now added to Extended Data Fig. 9. The Reporting Summary has been updated (the original uncorrected Reporting Summary is provided as the Supplementary Information to this Corrigendum). The original Letter has been corrected online.

Supplementary Information is available in the online version of this Corrigendum.

CORRECTIONS & AMENDMENTS

ERRATUM

doi:10.1038/nature26162

Erratum: Asparagine bioavailability governs metastasis in a model of breast cancer

Simon R. V. Knott, Elvin Wagenblast, Showkhin Khan, Sun Y. Kim, Mar Soto, Michel Wagner, Marc-Olivier Turgeon, Lisa Fish, Nicolas Erard, Annika L. Gable, Ashley R. Maceli, Steffen Dickopf, Evangelia K. Papachristou, Clive S. D'Santos, Lisa A. Carey, John E. Wilkinson, J. Chuck Harrell, Charles M. Perou, Hani Goodarzi, George Poulgiannis & Gregory J. Hannon

Nature **554**, 378–381 (2018); doi:10.1038/nature25465

In this Letter, several errors were inadvertently introduced during the production process. In Fig. 3d, the blue bars should be labelled 'L-asparaginase' rather than 'L-asparagine'. In the main text, "These were collected onto Matrigel..." should be "These were placed onto Matrigel...", and in the Methods, 'corresponding proteins' should be 'corresponding peptides' in the text: "...proteins were removed from subsequent analysis if fewer than 5 corresponding peptides were identified in any sample." The legend to Extended Data Fig. 6b should state 'Volumes of orthotopic..' rather 'Volumes of orthotropic...'. Finally, in the HTML the legend to Supplementary Table 4 was repeated for Supplementary Table 5. The legend to Supplementary Table 5 should have stated: 'Amino acid composition of serum with and without L-asparaginase treatment. 4T1-T cells harbouring the non-targeting *Renilla* shRNA were injected into immunocompromised mice. Five mice each were either injected with 60 U L-asparaginase or PBS 5 days per week. After blood collection and serum isolation, free amino acids were quantified using High Performance Liquid Chromatography (HPLC) and a fluorometric detector.' These errors have all been corrected online.

CORRECTIONS & AMENDMENTS

ERRATUM

doi:10.1038/nature26163

Erratum: Evolutionary history resolves global organization of root functional traits

Zeqing Ma, Dali Guo, Xingliang Xu, Mingzhen Lu,
Richard D. Bardgett, David M. Eissenstat, M. Luke McCormack
& Lars O. Hedin

Nature **555**, 94–97 (2018); doi:10.1038/nature25783

In this Letter, owing to an error during the production process, only author L.O.H. was listed as a corresponding author, instead of both D.G. (guodl@igsnrr.ac.cn) and L.O.H. (lhedin@princeton.edu). This has been corrected online.

Improving outcomes in congenital cataract

ARISING FROM H. Lin *et al.* *Nature* **531**, 323–328 (2016); doi:10.1038/nature17181

Lens regeneration after cataract surgery in infants is a clinical phenomenon with which paediatric ophthalmologists battle. Lin *et al.*¹ reported a novel surgical technique that aimed to convert this post-operative complication into an alternative therapy for children aged under 24 months. However, the early outcomes reported in the experimental group fall short of outcomes achievable through conventional treatment in this population. As the authors offer no comment or explanation for this discrepancy, this new approach cannot be considered effective or safe for affected children. There is a Reply to this Comment by Liu, Y. *et al.* *Nature* **556**, <https://dx.doi.org/10.1038/nature26150> (2018).

Regeneration of residual lens cells following surgical removal of congenital cataract can result in re-opacification, which needs to be treated by further intraocular surgery, necessitating repeated general anaesthetics during a sensitive period of neurodevelopment. The adaptation of this regenerative process by Lin *et al.*¹, in which a novel surgical method of cataract removal preserves endogenous lens cells, enabling functional lens regeneration, may eventually lead to the development of treatments for degenerative disease. However, issues relating to adult cataract (an age-related degenerative process) and congenital and infantile cataract are conflated in their report. Cataract is virtually universal in older age, making cataract surgery one of the most common surgical procedures, with enviably excellent visual outcomes. By contrast, infantile cataract is uncommon, affecting 3–15 per 10,000 children worldwide and, by definition, present from birth or early infancy². We now understand that mutations within the genes responsible for the production or orchestration of the lens epithelial progenitor/stem cells (LECs) are responsible for the majority of bilateral congenital or infantile cataract, even in cases where there is no family history³. Thus the treatment approach adapted by Lin *et al.*¹, which relies on regeneration of lens stem cells without addressing the persisting underlying genetic defect, cannot be definitive. Children treated using this technique may require further surgery but Lin *et al.* do not acknowledge this in the article.

The rationale for the trial reported by Lin *et al.*¹ is the need to address adverse outcomes associated with the use of artificial intraocular lenses, which are implanted in some children to replace the focusing power of the removed cataractous lens⁴. However, surgery with intraocular lens implantation was not the ‘control’ standard approach evaluated within the report. Equivalence in vision outcomes between their intervention and control groups was reported, but these should also have been assessed against the extant benchmark. Outcomes in infantile cataract have improved substantially over the past few decades, largely owing to the application of basic neuroscientific understanding of sensitive and critical periods in visual neurodevelopment. Cataract-related childhood visual impairment is largely due to bilateral stimulus deprivation amblyopia: the failure to restore a normal trajectory of visual neurodevelopment during a brief and finite window of opportunity. This critical window closes at some point during the first six months of life. Thus, younger age at surgery is the strongest predictor of better visual outcome and, in cataract present from birth, the window for intervention is conventionally considered to be the first six to eight weeks of life. Hence whole-population newborn screening programmes exist in many countries to ensure early diagnosis of congenital cataract and prompt referral for specialist treatment. In settings where such programmes do not yet exist, late diagnosis and treatment, resulting in irreversible amblyopia, drives poor visual outcomes. We have previously, on behalf of the British Isles Congenital Cataract Interest Group, reported outcomes within a contemporaneous, nationally

representative cohort of children undergoing surgery in the British Isles for congenital and infantile cataract in the first two years of life (IoLunder2 study)⁴. These outcomes are comparable to those found in other contemporary reports⁵ and are more than twofold better than those reported by Lin *et al.* in either their experimental or control groups. Indeed, the mean acuity achieved in their trial is the threshold for legal definition of blindness, an outcome that would lead most ophthalmologists, and probably most parents, to question the value of this new proposed intervention.

Effective treatment for congenital cataract requires, alongside surgery, post-operative management of the loss of refractive (focusing) power through removal of the natural lens. Failure to appropriately manage this results in dense amblyopia. As the method described in Lin *et al.* involves an 8-month post-operative period of lens regeneration with consequent partially obscured and poorly focused vision, the inevitable resultant amblyopia may explain the poor visual outcomes. The authors offer no other explanation, and do not describe how the rapidly changing and highly defocused refractive state (moving through 18 diopter units of refractive error in 8 months) was managed. Had the report adhered to the international reporting standard of CONSORT⁶ it might have been possible to assess the quality (internal validity) and generalizability (external validity) of the trial. For example, it is necessary to know: whether the control and intervention groups were equivalent with respect to baseline clinical characteristics (particularly age at surgery); how randomization was undertaken; the power calculation and the primary outcome and secondary outcomes on which this was based; how clustering by surgeon was addressed; and how clustering and correlation of outcome data were addressed in the analysis, given that both eyes of each subject were treated and analysed. The authors have described their study as a phase 1 trial, but the aim of such an investigation is to assess adverse outcomes of treatment, such as uncorrected high or irregular refractive outcome, which will drive the visual system towards amblyopia and resultant visual impairment.

As the paper stands, it is not possible to agree with its principal conclusion that it provides evidence supporting the superiority of the novel treatment. A tempered report, clearly articulating the limitations of the approach with respect to outcome and permanency of effect, would have avoided giving the false impression that the new approach can be expected to supersede current treatment practices.

A.L.S. and J.R. are supported by the National Institute for Health Research (NIHR) Biomedical Research Centres at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, and at the UCL Institute of Child Health/Great Ormond Street Hospital for children. A.L.S. is supported by fellowships from the Ulverscroft Vision Research Group and the Academy of Medical Sciences.

Ameenat Lola Solebo^{1,2,3,4}, Christopher J. Hammond⁵ & Jugnoo S. Rahi^{1,2,3,4}

¹Lifecourse Epidemiology and Biostatistics Section, Population, Policy and Practice Programme, Institute of Child Health, University College London, London, UK.

email: j.rahi@ucl.ac.uk

²Great Ormond Street Hospital/Institute of Child Health NIHR Biomedical Research Centre, London, UK.

³Moorfields Eye Hospital and Institute of Ophthalmology NIHR Biomedical Research Centre, London, UK.

⁴Ulverscroft Vision Research Group, University College London Institute of Child Health, London, UK.

⁵King's College Hospital London, Academic Section of Ophthalmology, School of Life Course Sciences, Faculty of Life Sciences and Medicine, St Thomas's Hospital Campus, London, UK.

Received 13 April 2016; accepted 23 January 2018.

1. Lin, H. *et al.* Lens regeneration using endogenous stem cells with gain of visual function. *Nature* **531**, 323–328 (2016).
2. Solebo, A. L. in *Congenital Cataract: A Concise Guide to Diagnosis and Management* (eds Lloyd, I. C. *et al.*) (Springer, 2016).
3. Gillespie, R. L. *et al.* Personalized diagnosis and management of congenital cataract by next-generation sequencing. *Ophthalmology* **121**, 2124–2137 (2014).

4. Solebo, A. L., Russell-Eggitt, I., Cumberland, P. M. & Rahi, J. S. Risks and outcomes associated with primary intraocular lens implantation in children under 2 years of age: the IoLunder2 cohort study. *Br. J. Ophthalmol.* **99**, 1471–1476 (2015).
5. Ye, H. H., Deng, D. M., Qian, Y. Y., Lin, Z. & Chen, W. R. Long-term visual outcome of dense bilateral congenital cataract. *Chin. Med. J. (Engl.)* **120**, 1494–1497 (2007).
6. Schulz, K. F., Altman, D. G. & Moher, D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med.* **7**, e1000251 (2010).

Author Contributions All authors contributed equally to all aspects of authorship.

Competing Financial Interests Declared none.

doi:10.1038/nature26148

Lens regeneration in children

ARISING FROM H. Lin *et al.* *Nature* **531**, 323–328 (2016); doi:10.1038/nature17181

Congenital cataracts are the primary cause of treatable childhood blindness worldwide, affecting about four infants per 10,000 live births¹. Current surgical techniques have helped thousands of patients, but the limitations of these techniques^{2–4} led Lin *et al.*⁵ to report an alternative approach that they claim leads to a regenerated lens with refractive capacity. We have concerns regarding features of the presented data and the conclusions reached in the Article. This has implications for our patients who ask for such surgical intervention. There is a Reply to this Comment by Liu, Y. *et al.* *Nature* **556**, <http://dx.doi.org/10.1038/nature26150> (2018).

Embryologically, the lens of the eye is derived from the surface ectoderm^{6,7} and shares many properties of ectodermal tissue. Lens cortex continues to be produced throughout life, and lens size continues to increase owing to progressive elongation of slowly dividing lens epithelial cells (LECs) into fibre cells that lose their nuclei^{8–13}. LECs are capable of unwanted proliferation in aged eyes; ophthalmologists observe this routinely after cataract extraction, when LECs proliferate and form semi-transparent scar-like tissue such as Soemmerring (Sömmerring) rings and lens pearls¹⁴. Lens transparency is maintained as the result of highly organized fibre cell packing with extracellular spaces that are narrower than the wavelength of light. The authors claim that if they perform a cataract extraction in a way that reduces the injury to the lens epithelium, the normal natural epithelial healing response will lead to a new biconvex, transparent lens with flexibility (accommodative power). However, the authors concede in their Reply to this BCA that with their intervention it is unreasonable to expect a completely normal regenerated lens. Without a completely normal lens, including clarity, normal shape, and normal size, one would predict that normal visual function would not develop and children's visual development would suffer from amblyopia. The lack of clarity and normal structure of the regenerated lentoids is highlighted for example in figure 3b of Lin *et al.*⁵, where the image of a rabbit lens seven weeks after surgery shows an optically irregularly shaped lens, not the biconvex shape of a normal lens⁶. In addition, figure 3c and f of ref. 5 shows images of rabbit lenses with posterior subcapsular cataracts in the visual axis that would decrease vision. Furthermore, histopathology was not presented in the rabbit regenerated lenses at maximum axial length, and thus extended data figure 7a of Lin *et al.*⁵ is insufficient to assess the quality and normality of lens regeneration. To demonstrate a normal reconstituted lens, it is necessary to provide sagittal lens sections and lens measurements that encompass the nodal point of the eye. The lens imperfections

demonstrated in figure 4a of ref. 5 are multiple, not single including the visual axis, and represent more than just the loss of LECs and scarring at the site of capsulorhexis. Such changes are by definition cataractous. The images of postoperative human eyes in figure 5f of ref. 5 appear to show hazy (that is, cataractous) lenses. More importantly, extended data figure 8e of ref. 5 shows no regenerated lens six months after lens surgery. Instead, a single light reflex is shown, which is likely to be from the lens capsule alone with no lens cortex and no evidence of a clear, biconvex crystalline lens that would be required for normal vision (compare the single lens capsule slit lens reflex of extended data figure 8e of ref. 5 with the normal lens reflex in figure 3b).

In the clinical setting, a clear view by indirect ophthalmoscopy (as shown in extended data figure 8c, d of ref. 5) in no way signifies that the lens casts a clear and focused image on the retina. It is easy for indirect ophthalmoscopy to get clear views through lenses with legally blinding cataracts. There is no information about the authors' management of the children's refractive status. Importantly, there remains the distinct possibility that the proposed surgical technique may cause inferior visual outcomes. Comparing visual acuity measures in studies of children can be inherently difficult because of different clinical characteristics and the use of different measurement techniques. Nevertheless, the visual outcomes reported in both the experimental and conventionally treated groups were unacceptably poor. Six months after surgery, the mean visual acuity was equivalent to 20/200 (6/60), whereas other investigators have reported that 60% of children with bilateral infantile cataracts achieve an acuity of 20/40 or better after standard ophthalmic care, and almost all achieve better acuity than the mean visual acuity reported by Lin *et al.*⁵. We would welcome a graphic, detailed distribution of the visual acuity results from the authors' patients (rather than only means or summarized data). Follow-up data covering more than one year represents another important clinical parameter. Finally, as the authors state that they did not exclude all genetic causes of congenital cataracts, and that they did not perform any gene editing or manipulation of the lens epithelial cells, we question why the proliferating LECs, presumably with genetic variants causing the original congenital cataracts, did not recapitulate the original lens opacity. One cannot just assume that germline mutations would constitute a minority of their cases without having data from their patients to support such a claim.

Although the current surgical approaches for congenital cataracts have adverse events and several limitations that have been identified in controlled long-term studies, Lin *et al.*⁵ do not provide sufficient short-term or long-term supportive evidence to support their statement that

BRIEF COMMUNICATIONS ARISING

their approach may afford an infant eye a good chance of an improved visual outcome. The study is limited by varied follow-up, poor visual acuities, insufficient details regarding adverse events, and a lack of information about the need for additional surgeries. These limitations should alert us to the need to proceed cautiously. Using such an unproven experimental therapy in both eyes of a patient is not appropriate.

We all support innovation and disruptive technology. Yet, at the same time, it is necessary to pay careful attention to details including thorough documentation and achievement of long-term milestones to support the conclusions. One should be even more vigilant when proposing a new therapy for our youngest patients, who represent the most vulnerable population. It is important to report both study results and limitations in a clear and balanced fashion, particularly to parents who desperately want the best possible outcome for their infants.

Demetrios G. Vavvas^{1,2}, Thaddeus P. Dryja³, M. Edward Wilson⁴, Timothy W. Olsen⁵, Ankoor Shah^{1,6}, Ula Jurkunas^{1,2}, Roberto Pineda^{1,2}, Vasiliki Poulaki^{1,7}, Sotiria Palioura⁸, Peter Veldman⁹, Javier Moreno-Montañés¹⁰, Maria D. Pinazo-Duran¹¹, José Carlos Pastor¹², Miltiadis Tsilimbaris¹³, Douglas Rhee¹⁴, Kathryn Colby⁹, David G. Hunter^{1,6}, Solon Thanos¹⁵, Taiji Sakamoto¹⁶, Louis R. Pasquale^{1,2}, Joan W. Miller^{1,2}, Deborah VanderVeen^{1,6} & Scott R. Lambert¹⁷

¹Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, USA.

²Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA.

email: Demetrios_Vavvas@meei.harvard.edu; Louis_Pasquale@meei.harvard.edu; Joan_Miller@meei.harvard.edu

³Ocular Pathology Service, Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, USA.

⁴Albert Florens Storm Eye Institute, Medical University of South Carolina, Charleston, South Carolina, USA.

email: wilsonme@musc.edu

⁵Department of Ophthalmology, Mayo Clinic, Rochester, Minnesota, USA.

email: Olsen.Timothy@mayo.edu

⁶Department of Ophthalmology, Boston Children's Hospital, Boston, Massachusetts, USA.

email: Deborah.Vanderveen@childrens.harvard.edu

⁷Department of Ophthalmology, Veterans Affairs Hospital, Boston University, Boston, Massachusetts, USA.

⁸Department of Ophthalmology, Bascom Palmer Eye Institute, University of Miami, Miami, Florida, USA.

⁹Department of Ophthalmology, University of Chicago, Chicago, Illinois, USA.

¹⁰Department of Ophthalmology, University of Navarra, Pamplona, Spain.

¹¹Ophthalmic Research Unit "Santiago Grisolia" and Cellular and

Molecular Ophthalmobiology Group at the Department of Ophthalmology, University of Valencia, Valencia, Spain.

¹²Department of Ophthalmology, Hospital Clinico Universitario and IOBA (Eye Institute) University of Valladolid, Valladolid, Spain.

email: pastor@ioba.med.uva.es

¹³Department of Ophthalmology, University of Crete, Crete, Greece.

¹⁴Department of Ophthalmology, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA.

¹⁵Institute of Experimental Ophthalmology, Westfalian Wilhelms-University of Münster, Albert-Schweitzer Campus 1, D15, 48149 Münster, Germany.

¹⁶Department of Ophthalmology, Kagoshima University, Kagoshima, Japan.

¹⁷Department of Ophthalmology, Stanford University School of Medicine, Stanford, California, USA.

email: scott.lambert@emory.edu

Received 13 April 2016; accepted 23 January 2018.

1. Wu, X., Long, E., Lin, H. & Liu, Y. Prevalence and epidemiological characteristics of congenital cataract: a systematic review and meta-analysis. *Sci. Rep.* **6**, 28564 (2016).
2. Lambert, S. R. *et al.* Is there a latent period for the surgical treatment of children with dense bilateral congenital cataracts? *J. AAPOS* **10**, 30–36 (2006).
3. Young, M. P., Heidary, G. & VanderVeen, D. K. Relationship between the timing of cataract surgery and development of nystagmus in patients with bilateral infantile cataracts. *J. AAPOS* **16**, 554–557 (2012).
4. Sukhija, J., Ram, J., Gupta, N., Sawhney, A. & Kaur, S. Long-term results after primary intraocular lens implantation in children operated less than 2 years of age for congenital cataract. *Indian J. Ophthalmol.* **62**, 1132–1135 (2014).
5. Lin, H. *et al.* Lens regeneration using endogenous stem cells with gain of visual function. *Nature* **531**, 323–328 (2016); corrigendum **541**, 558 (2017).
6. Cvekl, A. & Ashery-Padan, R. The cellular and molecular mechanisms of vertebrate lens development. *Development* **141**, 4432–4447 (2014).
7. Piatigorsky, J. Lens differentiation in vertebrates. A review of cellular and molecular features. *Differentiation* **19**, 134–153 (1981).
8. Augusteyn, R. C. On the growth and internal structure of the human lens. *Exp. Eye Res.* **90**, 643–654 (2010).
9. Hanna, C. & O'Brien, J. E. Cell production and migration in the epithelial layer of the lens. *Arch. Ophthalmol.* **66**, 103–107 (1961).
10. Harding, C. V., Hughes, W. L., Bond, V. P. & Schork, P. Autoradiographic localization of tritiated thymidine in wholemount preparations of lens epithelium. *Arch. Ophthalmol.* **63**, 58–65 (1960).
11. Mikulich, A. G. & Young, R. W. Cell proliferation and displacement in the lens epithelium of young rats injected with tritiated thymidine. *Invest. Ophthalmol.* **2**, 344–354 (1963).
12. Šikić, H., Shi, Y., Lubura, S. & Bassnett, S. A stochastic model of eye lens growth. *J. Theor. Biol.* **376**, 15–31 (2015).
13. Smith, P. Diseases of crystalline lens and capsule. 1. On the growth of the crystalline lens. *Trans. Ophthalmol. Soc. UK* **3**, 79–99 (1883).
14. Miyake, K., Ota, I., Miyake, S. & Horiguchi, M. Liquefied aftercataract: a complication of continuous curvilinear capsulorhexis and intraocular lens implantation in the lens capsule. *Am. J. Ophthalmol.* **125**, 429–435 (1998).

Author Contributions D.G.V., T.P.D., M.E.W., T.W.O., L.R.P., J.W.M., D.V. and S.R.L. conceived, wrote and edited the manuscript. A.S., U.J., R.P., V.P., S.P., P.V., J.M.-M., M.D.P.-D., J.C.P., M.T., D.R., K.C., D.G.H., S.T. and T.S. reviewed and edited the manuscript.

Competing Financial Interests Declared none.

doi:10.1038/nature26149

Liu *et al.* reply

Replying to: D. G. Vavvas *et al.* *Nature* **556**, <http://dx.doi.org/10.1038/nature26149> (2018); A. L. Solebo, C. J. Hammond & J. S. Rahi *Nature* **556**, <http://dx.doi.org/10.1038/nature26148> (2018)

In the accompanying Comments^{1,2}, Vavvas *et al.* describe issues regarding the normality of lens regeneration and Solebo *et al.* describe improving outcomes of congenital cataract. We welcome these Comments on our paper.

In our Article³, we did not claim that a completely normal lens was regenerated. Instead, our hypothesis was that current surgical methods for paediatric cataract, namely anterior continuous curvilinear capsulorhexis, may impair the integrity of lens epithelial stem cells (LECs).

BRIEF COMMUNICATIONS ARISING

Vavvas *et al.*¹ highlighted several figures as ‘cataractous’. While these lenses were mostly clear (particularly in the visual axis), we did not claim that they were completely normal and explicitly acknowledged their imperfections, which mainly reflected loss of LECs, mild peripheral scarring at the capsulorhexis site, and anterior–posterior capsule adhesions that resolve over time. The histopathology sections interpreted by Vavvas *et al.* as showing small or irregular lenses¹ were intentionally cut offset from the axial centre of the lens to minimize disruption to the lens cortex. Furthermore, dissected lenses often shrink upon alcohol dehydration. Thus, irregularities in size or shape were essentially fixation artefacts. Vavvas *et al.*¹ also comment on the lack of quality office photographs, although infants usually do not cooperate with quality office photography, and this was also not done in the recent NIH Infant Aphakia Treatment Study (IATS)⁴.

Regarding concerns about amblyopia, all patients in our study underwent monthly post-operative correction of refractive error with either spectacles or contact lenses to maintain an appropriate refractive state consistent with age.

Regarding visual outcomes, the studies^{5,6} cited by Vavvas *et al.*¹ are not comparable to ours for several reasons. First, the mean age was higher (5.3 years and 10.2 years compared with under 2 years in our study), which may have confounded visual outcomes, as the IATS noted that follow-up length can affect visual acuity comparisons⁴. Second, the evaluation of visual acuity differed (Teller Acuity Cards for visual resolution³ versus Snellen Visual Acuity for visual recognition^{5,6}). Although Teller cards can be roughly translatable to Snellen equivalents, accuracy and false comparison concerns usually preclude this. To illustrate, the IATS reported logMAR grating acuities at 1 year of age and not Snellen equivalents⁴. Third, only 60% of children had 20/40 or better in ref. 5, and only 70% were able to complete Snellen visual acuity testing in ref. 6. The values reported in those studies therefore likely reflect bias of the data towards a seemingly better outcome.

Similarly, Solebo *et al.*² also refer to a group of children operated on via primary intraocular lens implantation at a median of 8 months of age⁷. The description provided in ref. 7 leads us to believe that these children may have a different form of cataract (developmental) than the one exhibited by our group of patients (congenital), making direct comparisons of outcome difficult and inappropriate owing to the variation in impact on amblyopia.

In commenting on our outcomes, Solebo *et al.*² did not account for normal age-related changes in visual acuity. During development, 20/200 is close to normal vision for six months of age, and 20/50–20/80 is close to normal vision for one year of age, depending upon the method of testing used. The visual acuity levels in our population were appropriate and not inferior to their results if age is appropriately considered. In addition, an error in vision reported in our initial paper was corrected in a subsequent Corrigendum⁸. Therefore, their comments on our visual acuity results are now out of context and inaccurate.

Both Vavvas *et al.*¹ and Solebo *et al.*² mention concern of recurrent cataracts in children with underlying germline mutations. We agree that in these patients, cataracts would be expected to recur eventually. However, we excluded infants with a family history of ocular disease and were not aware of any inherited mutations among the children in our study. Note that lens removal and placement of an intraocular lens implant (IOL) also does not address any underlying genetic mutations. All the regenerated lenses in our study were all initially transparent. Only two lenses became cloudy again after one year and required further surgery. However, those patients had a relatively clear lens during a critical period in visual development, thus avoiding high risk of amblyopia.

New techniques in medicine deserve careful consideration, as well as established reproducibility, before adoption. The intent of this minimally invasive surgical method for paediatric cataracts was precisely to “First do no harm;” that is, to allow regeneration of a functional lens with greater transparency of the visual axis, while avoiding the side effects and complications associated with current methods. The IATS reported an 81% incidence of adverse events when IOLs were implanted in babies, and 72% of their IOL group required additional surgery under general anaesthesia versus 21% of the aphakic group⁹. Our new approach may afford the infant eye a better chance for improved ocular development and visual outcome, without the need for a more invasive procedure with increased risk of ocular and possibly neural complications. We are currently conducting a longer-term follow up study to gain additional data about long-term safety and efficacy. Generation of a completely normal and functional lens is our ultimate goal; this will require the creation of a conducive environment and scaffold for LES proliferation and differentiation.

The author list of this Reply comprises those authors involved in clinical investigation of lens regeneration. Those authors of the original paper who were involved in animal and cell culture experiments and did not contribute to this response are not listed here.

Yizhi Liu¹, David Granet², Haotian Lin¹, Sally Baxter², Hong Ouyang¹, Jie Zhu², Shan Huang¹, Zhenzhen Liu¹, Xiaokang Wu³, Fangbing Yan³, Xialin Liu¹, Lixia Luo¹, Christopher Heichel², Meixia Zhang³, Wenjia Cai^{1,2}, Richard L. Maas⁴ & Kang Zhang^{1,2,3,5,6}

¹State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China.

email: kang.zhang@gmail.com

²Shiley Eye Institute, University of California, San Diego, La Jolla, California 92093, USA.

³Molecular Medicine Research Center, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Sichuan 610041, China.

⁴Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.

⁵Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou 510060, China.

⁶Veterans Administration Healthcare System, San Diego, California 92093, USA.

1. Vavvas, D. G. *et al.* Lens regeneration in children. **556**, *Nature* <http://doi.org/10.1038/nature26149> (2018).
2. Solebo, A. L. *et al.* Improving outcomes in congenital cataract. *Nature* **556**, <http://doi.org/10.1038/nature26148> (2018).
3. Lin, H. *et al.* Lens regeneration using endogenous stem cells with gain of visual function. *Nature* **531**, 323–328 (2016).
4. Drews-Botsch, C. D., Celano, M., Kruger, S & Hartmann, E. E. Adherence to occlusion therapy in the first six months of follow-up and visual acuity among participants in the Infant Aphakia Treatment Study (IATS). *Invest. Ophthalmol. Vis. Sci.* **53**, 3368–3375 (2012).
5. Lambert, S. R. *et al.* Is there a latent period for the surgical treatment of children with dense bilateral congenital cataracts? *J. AAPOS* **10**, 30–36 (2006).
6. Young, M. P., Heidary, G. & VanderVeen, D. K. Relationship between the timing of cataract surgery and development of nystagmus in patients with bilateral infantile cataracts. *J. AAPOS* **16**, 554–557 (2012).
7. Solebo, A. L., Russell-Eggitt, I., Cumberland, P. M. & Rahi, J. S. Risks and outcomes associated with primary intraocular lens implantation in children under 2 years of age: the IoLunder2 cohort study. *Br. J. Ophthalmol.* **99**, 1471–1476 (2015).
8. Lin, H. *et al.* Corrigendum: Lens regeneration using endogenous stem cells with gain of visual function. *Nature* **541**, 558 (2017).
9. Infant Aphakia Treatment Study Group. Comparison of contact lens and intraocular lens correction of monocular aphakia during infancy: a randomized clinical trial of HOTV optotype acuity at age 4.5 years and clinical findings at age 5 years. *JAMA Ophthalmol.* **132**, 676–682 (2014).

doi:10.1038/nature26150

Investigating non-Joulian magnetostriction

ARISING FROM H. D. Chopra & M. Wuttig *Nature* **521**, 340–343 (2015); doi:10.1038/nature14459

Ferromagnetic materials change their shape under an applied magnetic field—a phenomenon known as magnetostriction. This phenomenon was first described for iron by Joule in 1842, and is generally believed to be volume-conserving. We therefore read with interest the Letter by Chopra and Wuttig¹, which reports that samples of Fe_{100-x}Ga_x (galfenol) crystals demonstrate “giant” non-volume-conserving (non-Joulian) magnetostriction, and embarked on an extensive study of crystals of similar compositions, dimensions and heat treatments, using strain gauges and capacitive dilatometry to measure magnetostriction for many different combinations of strain direction and applied magnetic field. In every case, we found that the volume was conserved within experimental error, and so we conclude that magnetostriction in galfenol can generally be regarded as Joulian.

The initial claim¹ was based on strain-gauge measurements in the plane of slow-cooled or quenched, disk-shaped crystals of Fe_{82.9}Ga_{17.1} and Fe_{73.9}Ga_{26.1}. These samples were 5 mm in diameter and 0.4–0.5 mm thick. A magnetic field of up to 3,000 Oe was applied in-plane. The crystals were found to expand in all, or almost all, of the directions that were tested. No measurement of strain was reported for the [001] direction perpendicular to the plane of the disks because it was assumed that “a negligible magnetization normal to the plane of the disk at comparable fields implies that no volume change occurs along [001].”¹ Strain-gauge data for the [001] direction of another slow-cooled Fe_{82.9}Ga_{17.1} crystal were published subsequently as an Addendum², in support of the original assertion.

In our attempt to verify the idea that the enhanced magnetostriction of galfenol is largely non-Joulian, we grew 12 different rod-, disk- or cuboid-shaped crystals with compositions of Fe₈₃Ga₁₇ or Fe₇₄Ga₂₆, heights (or thicknesses) ranging from 70 mm to 0.1 mm and diameters (or widths) ranging from 16 mm to 7 mm. Here we focus on the four of these crystals that were subjected to the same heat treatments as described in ref. 1 (annealed at 1,033 K for 30 min and then either quenched or slow-cooled at 10 K min⁻¹). We measured the magnetostriction (λ , in parts per million (p.p.m.)) using strain gauges or, for the disk-shaped crystals with thicknesses of 0.5 mm, by capacitive dilatometry³ in the [001] direction.

The most complete datasets were obtained for a single crystal of Fe₈₃Ga₁₇ with dimensions of 10.6 mm × 10.6 mm × 2.4 mm in the quenched and slow-cooled states. For this crystal we measured the saturation magnetostriction along the [100], [010], [001], [110] and [110] directions in a magnetic field large enough to saturate the magnetization applied along any one of these directions, yielding 17 independent measurements for the crystal in each state. Defining the magnetostriction components (in p.p.m.) for a given applied field (i) and measurement (j) direction as λ_{ij}^q and λ_{ij}^{sc} for the quenched and slow-cooled states, respectively, and using brackets ‘[...]’ to denote an array of these components, we find

$$\begin{aligned} [\lambda_{ij}^q] &= \begin{bmatrix} 161(155) & -80(-80) & -75(-75) \\ -85(-82) & 160(157) & -75(-75) \\ -80(-82) & -80(-80) & 155(162) \end{bmatrix}; \sum_j [\lambda_{ij}^q] = \begin{bmatrix} 6(0) \\ 0(0) \\ -5(0) \end{bmatrix} \\ [\lambda_{ij}^{sc}] &= \begin{bmatrix} 110(115) & -111(-110) & 4(-5) \\ -110(-110) & 113(115) & 2(-5) \\ -113(-110) & -113(-110) & 226(220) \end{bmatrix}; \sum_j [\lambda_{ij}^{sc}] = \begin{bmatrix} 3(0) \\ 5(0) \\ 0(0) \end{bmatrix} \end{aligned} \quad (1)$$

for $i, j \in \{[100], [010], [001]\}$ and

$$\begin{aligned} [\lambda_{ij}^q] &= \begin{bmatrix} 44(42) & 35(33) & -75(-75) \\ 35(33) & 43(42) & -76(-75) \\ -78(-81) & -80(-81) & 155(162) \end{bmatrix}; \sum_j [\lambda_{ij}^q] = \begin{bmatrix} 4(0) \\ 2(0) \\ -3(0) \end{bmatrix} \\ [\lambda_{ij}^{sc}] &= \begin{bmatrix} 12(8) & 3(-3) & -5(-5) \\ 2(-3) & 16(8) & -6(-5) \\ -105(-110) & -108(-110) & 225(220) \end{bmatrix}; \sum_j [\lambda_{ij}^{sc}] = \begin{bmatrix} 10(0) \\ 12(0) \\ 12(0) \end{bmatrix} \end{aligned} \quad (2)$$

for $i, j \in \{[110], [1\bar{1}0], [001]\}$. Fitting the measured data to minimize the discrepancy between the measured values and calculated values (see below) yields the numbers given in parentheses in equations (1) and (2).

The saturation magnetostriction λ_{ij}^s is the linear strain $\Delta l_{ij}/l_{ij}$ in an applied field that is large enough to saturate the magnetization. We compare our data with the results expected from the standard expression for Joulian magnetostriction of a crystal with cubic symmetry⁴:

$$\begin{aligned} \lambda_{ij}^s &= \frac{3}{2} \lambda_{100} \left(\alpha_x^2 \beta_x^2 + \alpha_y^2 \beta_y^2 + \alpha_z^2 \beta_z^2 - \frac{1}{3} \right) \\ &\quad + 3 \lambda_{111} (\alpha_x \alpha_y \beta_x \beta_y + \alpha_y \alpha_z \beta_y \beta_z + \alpha_z \alpha_x \beta_z \beta_x) \end{aligned} \quad (3)$$

where λ_{100} and λ_{111} are the two intrinsic saturation magnetostriction constants of the cubic material, and α and β are the cosines that define the magnetization direction (i) and the measurement direction (j), respectively; for example, when the field is applied along $i = [100]$ (the x direction) and the magnetostriction is measured along $j = [010]$ (the y direction), $\alpha_x = 1$, $\alpha_y = \alpha_z = 0$, $\beta_y = 1$ and $\beta_x = \beta_z = 0$. More general expressions for other lattice symmetries have been derived previously⁵.

We make no assumptions about the domain structure in the initial demagnetized state. Defining n_x , n_y and n_z as the fractions of domains oriented along the three principal directions, [100], [010] and [001], respectively, and assuming that the strains are additive, the predicted magnetostriction based on equation (3) is

$$[\lambda_{ij}^s] = \frac{3}{2} \lambda_{100} \begin{bmatrix} n_y + n_z & -n_y & -n_z \\ -n_x & n_x + n_z & -n_z \\ -n_x & -n_y & n_x + n_y \end{bmatrix} \quad (4)$$

for $i, j \in \{[100], [010], [001]\}$ and

$$\begin{aligned} [\lambda_{ij}^s] &= \frac{3}{4} \lambda_{100} \begin{bmatrix} n_z & n_z & -2n_z \\ n_z & n_z & -2n_z \\ -(n_x + n_y) & -(n_x + n_y) & 2(n_x + n_y) \end{bmatrix} \\ &\quad + \frac{3}{4} \lambda_{111} \begin{bmatrix} n_x + n_y & -(n_x + n_y) & 0 \\ -(n_x + n_y) & n_x + n_y & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (5)$$

for $i, j \in \{[110], [1\bar{1}0], [001]\}$. As in equations (1) and (2), the rows of $[\lambda_{ij}^s]$ in equations (4) and (5) correspond to the applied field direction (i) and the columns to the measurement directions (j).

Consider the first element in the array in equation (4) as an example, $i = j = [100] \equiv 1$. For an initially $\pm x$ -aligned domain, $\alpha_x = \pm 1$ and $\alpha_y = \alpha_z = 0$. Because we are measuring along the x direction, $\beta_x = 1$ and $\beta_y = \beta_z = 0$. Putting these numbers into equation (3), we find, $\lambda_{11}^s = \lambda_{100}$. But there is no difference in strain between the initial and final states when we flip the $\alpha_x = -1$ domains; it remains $\lambda_{11}^s = \lambda_{100}$. The shape of the sample is unchanged by the magnetization process of the x -aligned domains, so they do not contribute to this element of the magnetostriction. However, for an initially y -aligned domain, $\lambda_{11}^s = -\lambda_{100}/2$, because $\alpha_x = \alpha_z = 0$, $\alpha_y = \pm 1$, $\beta_x = 1$ and $\beta_y = \beta_z = 0$ when we measure along x . A field applied along x rotates the magnetization of the initially y -aligned domains to $\alpha_x = 1$ and $\alpha_y = \alpha_z = 0$, yielding $\lambda_{11}^s = \lambda_{100}$. The net magnetostriction contributed by y -aligned domains is therefore $\lambda_{100} - (-\lambda_{100}/2) = 3\lambda_{100}/2$. The same is true for an initially z -aligned domain: $\lambda_{11}^s = 3\lambda_{100}/2$. Consequently, $\lambda_{11}^s = 0 \times n_x + 3\lambda_{100}/2 \times n_y + 3\lambda_{100}/2 \times n_z = 3\lambda_{100}(n_y + n_z)/2$, as shown in equation (4). Similar arguments can be used to derive the other elements in equations (4) and (5).

The Joulian condition states that the sum of the elements of each row of $[\lambda_{ij}^s]$ is zero; this condition is fulfilled in equations (4) and (5). Furthermore, λ_{100} can be deduced from the elements in any column of either equation independently of the initial domain structure because $n_x + n_y + n_z = 1$. There are therefore three unknown parameters (n_x , n_y and λ_{100}) for the $[100]$, $[010]$ and $[001]$ directions and four (n_x , n_y , λ_{100} and λ_{111}) for the $[110]$, $[1\bar{1}0]$ and $[001]$ directions; but there are many more data in equations (1) and (2), in which the best fits to the data are given in parentheses. The magnetostriction is well fitted by the Joulian model with fit parameters $n_x = 35\%$, $n_y = 34\%$ (implying $n_z = 31\%$), $3\lambda_{100}/2 = 236 \pm 6$ p.p.m. and $3\lambda_{111} = 25 \pm 3$ p.p.m. for the quenched state of $\text{Fe}_{83}\text{Ga}_{17}$, and $n_x = 49\%$, $n_y = 49\%$ (implying $n_z = 2\%$), $3\lambda_{100}/2 = 225 \pm 4$ p.p.m. and $3\lambda_{111} = 23 \pm 8$ p.p.m. for the slow-cooled

Table 1 | Domain distributions and magnetostriction in disk-shaped $\text{Fe}_{100-x}\text{Ga}_x$ crystals

	Row sum, $\sum_j [\lambda_{ij}^{\text{q,sc}}]$ (p.p.m.)		n_x (%)	n_y (%)	n_z (%)	$3\lambda_{100}/2$ (p.p.m.)
	$i = [100]$	$i = [010]$				
$x = 17$ (quenched)	-5	-10	85	15	0	265 ± 7
$x = 17$ (slow-cooled)	13	5	40	40	20	251 ± 7
$x = 26$ (quenched)	-3	0	44	45	11	174 ± 3
$x = 26$ (slow-cooled)	10	10	38	34	28	160 ± 4

state. The small discrepancies between the measured and fitted values—of 2 p.p.m. and 4 p.p.m. for the quenched and slow-cooled states, respectively, when averaging the magnitudes of the differences over the 17 independent measurements ($(1/17) \sum_{ij} |\lambda_{ij}^{\text{q,sc}} - \lambda_{ij}^{\text{q,sc,fit}}|$)—could be related to the width of the domains relative to the width and position of the strain gauges or to the uniformity of the initial state. The average magnitude of the row sum is 3 p.p.m. for the quenched crystal and 7 p.p.m. for the slow-cooled one.

We also have nearly complete $\langle 100 \rangle$ data for quenched and slow-cooled, disk-shaped (diameter, 5 mm; thickness, 0.5 mm) crystals of $\text{Fe}_{83}\text{Ga}_{17}$ and $\text{Fe}_{74}\text{Ga}_{26}$. Only measurements of λ_{33} ($i = j = [001] \equiv 3$) are missing, because the thin disks tend to twist in the large perpendicular field that is needed for saturation. For this crystal we obtained eight independent measurements, for three fit parameters. These data are also in excellent agreement with the Joulian model in equation (3), with average discrepancies of 2–3 p.p.m. between the measured and calculated magnitudes of λ_{ij}^s . The results of the fit and the row sums ($\sum_j [\lambda_{ij}^{\text{q,sc}}]$) for fields applied along the $i = [100]$ and $i = [010]$ directions are summarized in Table 1. Separate determinations of λ_{100} obtained

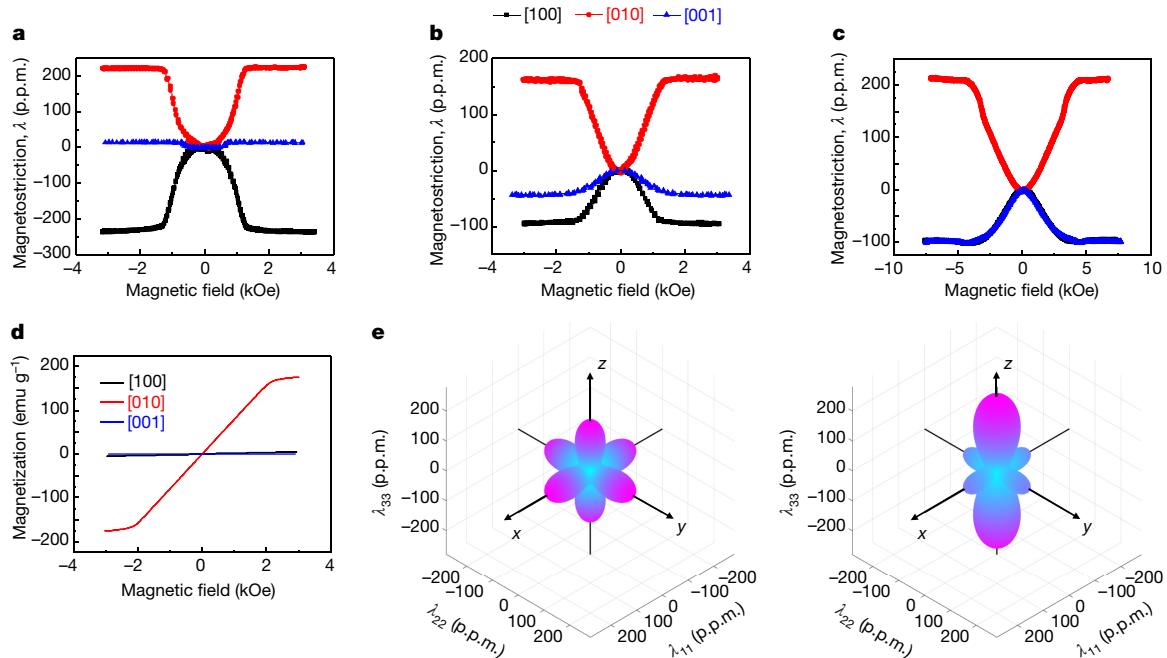


Figure 1 | Magnetostriction and magnetization of $\text{Fe}_{83}\text{Ga}_{17}$ single crystals. a, b, Magnetostriction λ for quenched (a) and slow-cooled (b) disk-shaped crystals (diameter, 5 mm; thickness, 0.5 mm) in an in-plane magnetic field applied along the $i = [010]$ direction, measured in three orthogonal directions: $j = [100]$ (black squares), $j = [010]$ (red circles) and $j = [001]$ (blue triangles). c, Magnetostriction measured similarly for an as-grown 10 mm \times 10 mm \times 10 mm cubic crystal. d, Magnetization curves for an as-grown disk-shaped crystal (diameter, 5 mm; thickness, 0.88 mm)

measured with the field along the $[010]$ direction. The saturation field in a–d depends on the shape of the sample. e, Magnetostriction calculated from equation (3), with $3\lambda_{100}/2 = 250$ p.p.m. and $3\lambda_{111} = 26$ p.p.m., when the measurement direction is parallel to the magnetization direction in saturation, for an isotropic domain distribution (top; $n_x = n_y = n_z = 1/3$) and an in-plane domain distribution (bottom; $n_x = n_y = 1/2$, $n_z = 0$). The radial coordinate represents the magnetostriction (in p.p.m.), while the angular coordinates indicate the direction of the saturation magnetization.

by summing the elements in the first or second columns of the array in equation (4), which are independent of the domain structure, agree with the values in Table 1 to within 5 p.p.m.

Our results suggest that the heat treatment modifies the initial domain structure of the crystals, which have the same dimensions as those studied in ref. 1, but that it has only a small effect on the magnetostriction for a given composition. Data for $\text{Fe}_{83}\text{Ga}_{17}$ in the quenched and slow-cooled states are shown in Fig. 1a, b; data for a cubic crystal with an isotropic initial domain distribution are shown in Fig. 1c for comparison. Our crystals exhibit zigzag-striped domains, with 90° or 180° domain walls, as described previously^{6–8}. We found no sign of the small periodic Landau closure domains that are evident in images of the surface of a $\text{Fe}_{73.9}\text{Ga}_{26.1}$ crystal in figure 3 of ref. 1.

The magnetization curves for our disk-shaped samples (Fig. 1d) resemble those reported in ref. 1, but the negligible magnetization normal to a disk magnetized in an in-plane field is accompanied by substantial perpendicular magnetostriction whenever z domains are present, represented by the $-(3/2)n_z\lambda_{100}$ terms in equation (4). Altogether, we have obtained 31 row sums from measurements of 12 different crystals, with an overall average of 6 ± 7 p.p.m. None of these row sums is representative of the sample in ref. 2, for which the average row sum is 134 p.p.m.

Finally, we repeated Joule's original 1846 experiment⁹, measuring the saturation volume magnetostriction of a rod-shaped $\text{Fe}_{83}\text{Ga}_{17}$ crystal (diameter, 16 mm; height, 70 mm) using the capillary liquid displacement method in alcohol in a field of 2 kOe. This method avoids the shortcomings of strain gauges. The crystal was measured as-grown and after quenching or slow-cooling. In all cases, the volume expansion was found to be less than 5 p.p.m.

In summary, we have found no evidence of giant non-Joulian magnetostriction in any of the galfenol crystals that we studied. We do not exclude the possibility of minor non-Joulian contributions associated with, for example, forced volume magnetostriction, auxetic behaviour or the 'Δ*E* effect' in the unsaturated state. However, we conclude that the large intrinsic magnetostriction of galfenol is essentially Joulian and that the findings of ref. 1 are not generalizable to this class of iron-based magnetostrictive materials.

Data Availability The data generated and analysed during this study are available from the corresponding author on reasonable request.

Yangkun He^{1,2}, Yongjun Han¹, P. Stamenov², B. Kundys³, J. M. D. Coey^{1,2}, Chengbao Jiang¹ & Huibin Xu¹

¹School of Materials Science and Engineering, Beihang University, 100191 Beijing, China.

email: jiangcb@buaa.edu.cn

²School of Physics and AMBER, Trinity College, Dublin 2, Ireland.

³PCMS, University of Strasbourg, 67034 Strasbourg, France.

Received 2 December 2015; accepted 9 January 2018.

1. Chopra, H. D. & Wuttig, M. Non-Joulian magnetostriction. *Nature* **521**, 340–343 (2015).
2. Chopra, H. D. & Wuttig, M. Addendum: Non-Joulian magnetostriction. *Nature* **538**, 416 (2016).
3. Kundys, B. *et al.* Three terminal capacitance technique for magnetostriction and thermal expansion measurements. *Rev. Sci. Instrum.* **75**, 2192 (2004).
4. Cullity, B. D. & Graham, C. D. *Introduction to Magnetic Materials* 2nd edn, Ch. 6, 245 (Wiley, 2009).
5. Lee, E. W. Magnetostriction and magnetomechanical effects. *Rep. Prog. Phys.* **18**, 184–229 (1955).
6. Mudivarthi, C. *et al.* Magnetic domain observations in Fe–Ga alloys. *J. Magn. Mater.* **322**, 2023–2026 (2010).
7. Asano, S. *et al.* Magnetic domain structure and magnetostriction of Fe–Ga alloy single crystal grown by the Czochralski method. *IEEE Magn. Lett.* **8**, 6101004 (2017).
8. He, Y. K., Coey, J. M. D., Shaefer, R. & Jiang, C. Determination of bulk domain structure and magnetization process in ferromagnetic bcc alloys: analysis of magnetostriction in $\text{Fe}_{83}\text{Ga}_{17}$ crystals. *Phys. Rev. Mater.* **2**, 014412 (2018).
9. Joule, J. P. XVII. On the effects of magnetism upon the dimensions of iron and steel bars. *Phil. Mag. J. Sci.* **30**, 76–87 (1847).

Author Contributions C.J., J.M.D.C. and H.X. designed the experiment. Y. Han and Y. He grew the single crystals. Y. He, P.S. and B.K. conducted the magnetostriction measurements. Y. He, J.M.D.C., P.S. and C.J. analysed the data and wrote the paper. All authors discussed the results.

Competing Financial Interests Declared none.

doi:10.1038/nature25780

ECOLOGY'S REMOTE-SENSING REVOLUTION

Satellite data, and the tools that ecologists use to analyse them, are more accessible and plentiful than ever.

ILLUSTRATION BY THE PROJECT TWINS



BY ROBERTA KWOK

When ecologist Nicholas Murray started digging into remote-sensing data for his PhD project, he had no idea how hard his task would be. Murray wanted to know why shorebirds that migrate through Asia were declining in number. Because the birds stopped in places that were difficult for Murray to access, such as North Korea and China, he turned to satellite data to evaluate

their habitat.

When Murray started the project in 2010 he guessed it would take a few months, but it ended up taking about a year. Murray first had to download metadata for about 5,500 publicly available US government satellite images to identify those of tidal wetlands taken during low tide along the Yellow Sea, which borders China and the Korean peninsula. He then wrote custom software code to classify land cover in a final set of 80 images. An algorithm to

distinguish water from land already existed, but he needed to make manual adjustments for each image. More than one-quarter of the wetland area had vanished between the 1980s and 2000s, Murray discovered. But the analysis wasn't easy. "Throughout that whole process, I was thinking, 'This is so difficult, it's unbelievable,'" recalls Murray, now at the University of New South Wales in Kensington, Australia.

Today, Murray's task would be much simpler. Numerous tools have been developed to ►

► access and analyse remote-sensing data, allowing ecologists to tackle large-scale conservation problems more easily. Government agencies, open-source developers and commercial firms are offering everything from point-and-click interfaces to command-line-driven software. “I think we’re at the best time possible to be doing it,” Murray says of analysing satellite data for ecology research. “It’s getting so accessible.”

‘Remote sensing’ encompasses a suite of techniques for observing something without touching it. The term usually refers to collecting data about Earth from space or from airborne platforms by measuring energy reflected or emitted at various wavelengths. Researchers can use these data to infer, for example, the level of deforestation. “We’ve seen a real explosion in the use of satellite data,” says Allison Leidner, a contract senior support scientist at NASA’s Biological Diversity research programme in Washington DC.

Landsat data, gathered by NASA and the US Geological Survey (USGS), extend back to the 1970s and enable the study of planetary change over many decades. NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS) instruments, launched in 1999 and 2002, measure reflected solar radiation and emitted radiation, and the data are automatically converted into ecologist-friendly parameters such as vegetation greenness. And Europe’s Sentinel satellites, which monitor the land, ocean and atmosphere, have been providing data since 2014.

Users can browse free government data sets at online portals such as NASA’s Earthdata Search, EarthExplorer from the USGS and the European Space Agency’s Copernicus Open Access Hub. Earth data are typically divided into sections called ‘scenes’ or ‘tiles’ — snapshots of energy of varying wavelengths reflected from that area. But to obtain higher spatial and temporal resolution, researchers might want to consider commercial options.

The Dove satellites operated by Planet in San Francisco, California, for example, gather global data at a resolution of 3.7 metres — sharp enough to distinguish individual large trees — about once a day. The Sentinel-2 satellites, by contrast, which are among the highest-resolution government satellites with free and open data, have 10-metre pixels and sample each spot every 5 days. University researchers can apply for free access to 10,000 square kilometres of Planet’s satellite data per month through the firm’s Education and Research programme. Similarly, academic environmental researchers can apply for free access to data from sub-metre-resolution satellites operated by DigitalGlobe in Westminster, Colorado, through the non-profit organization DigitalGlobe Foundation.

USER-FRIENDLY ACCESS

Data sets can be unwieldy, however. Kyla Dahlin, an ecogeographer at Michigan State University

in East Lansing, notes that 30 years of data collected for one Landsat scene could exceed 1.5 terabytes “for an area that’s smaller than Michigan”. Visualization software for remote-sensing data might not function well with certain file formats, and although these files can be converted into an easier-to-use format, that step adds another obstacle, says Cindy Schmidt, associate programme manager of Ecological Forecasting Applications at the NASA Ames Research Center in Moffett Field, California. Inexperienced users “just want to throw in the towel sometimes”, she says. “They don’t have time to deal with that kind of stuff.”

Free and commercial resources are available, however. In 2017, Murray’s team released a free online tool called Remap, which enables users to generate maps from remote-sensing data. Users train the software to classify land-cover types, such as forest or wetlands, by uploading geo-referenced data or identifying pixels on the basis of fieldwork or their knowledge. Remap then uses machine learning to classify the remaining pixels. As of March 2018, about 4,300 people from more than 100 countries have used Remap, Murray says. Another online tool, called Global Forest Watch, creates maps of deforestation patterns.

Dahlin recommends the online tool AppEARS (Application for Extracting and Exploring Analysis-Ready Samples), which allows users to grab data specific to their study site, instead of an entire tile or scene. “Imagine the archive is this big lake of data,” explains Tom Maersperger, project scientist at the NASA Land Processes Distributed Active Archive Center (DAAC) in Sioux Falls, South Dakota, which led the tool’s development. “We’re allowing people to come in with a syringe and suck up that little sample that they want.” Users can provide geographical coordinates, a time span and variables of interest — such as tree cover — and the software returns the data as a comma-separated-values (CSV) file.

Similarly, the US Oak Ridge National Laboratory DAAC, in Tennessee, has tools to provide for example a time series of greenness for a study site as a spreadsheet and graph, or processed data, such as inferred forest disturbance. Ecologists can then analyse links between vegetation and other variables such as animal populations. One team, for example, studied the Andaman Islands off the Indian coast and found that vegetation degraded more quickly in areas where elephants and spotted deer had been introduced.

For ecologists who want to write their own analysis software, but avoid the hassle of downloading satellite data, Google Earth Engine is a popular choice. Google has already downloaded satellite data sets onto its servers, and researchers can access them in the cloud for free through Google’s JavaScript and Python programming interfaces. This service allows researchers to perform large-scale analyses much faster than they could on their local computers.

Murray, for instance, leveraged that processing power to map global intertidal zones over time. Because it used more than 700,000 satellite images, the analysis would have taken years on a single computer — but it took less than a week on Google Earth Engine. The tool has “revolutionized the sorts of remote sensing questions I can ask”, Murray says.

Google says that users need not worry that it will claim ownership over their intellectual property (IP), such as code and scientific results. “Our terms of service make it clear that your IP is your IP and we make no claims on it,” says Noel Gorelick, an engineer at Google in Zürich, Switzerland, who co-developed Google Earth Engine. Still, Martin Wegmann, a remote-sensing researcher at the University of Würzburg in Germany, prefers to download satellite data and run his code locally. Because his analyses are relatively small in scale or coarse in resolution, performance is not an issue, he says.

Other cloud-computing options include the Centre for Environmental Data Analysis, run by the Science and Technology Facilities Council in Harwell, UK; Copernicus Data and Information Access Services, funded by the European Commission and scheduled to go live in June; and DigitalGlobe’s GBDX platform.

OPEN-SOURCE OPTIONS

Whichever platform they choose, researchers typically write custom code to drive data analysis, often in the programming language R. Wegmann and his colleagues are developing an R package called *getSpatialData*, which will allow users to download satellite data without using a browser interface. His team also developed the *RStoolbox* package, which includes different algorithms for computing vegetation measures so that users do not have to calculate specific formulae individually.

Researchers can also use commercial desktop analysis and visualization packages such as ENVI from Harris in Melbourne, Florida; ERDAS IMAGINE from Hexagon Geospatial in Madison, Alabama; ArcGIS from Esri in Redlands, California, as well as free, open-source alternatives such as QGIS.

Because using these tools can involve steep learning curves, Anita Graser, a geographic information scientist at the Austrian Institute of Technology in Vienna and a member of the QGIS project steering committee, advises beginners to take online classes. NASA’s Applied Remote Sensing Training programme offers webinars, and the agency gives workshops at ecology and conservation conferences.

The possibilities are enticing, but researchers must remember to stay grounded. “If you wanted to see how often a butterfly visits a nectar plant, you’re not going to pick that up on a satellite,” Leidner says. But for larger-scale problems, “it’s an incredibly powerful tool”. ■

Roberta Kwok is a freelance writer in Kirkland, Washington.

CAREERS

INTERNATIONAL STUDENTS Canada gains popularity **p.141**

LEADERSHIP Fewer top universities have female presidents than in 2017 **p.141**

NATUREJOBS For the latest career listings and advice www.naturejobs.com

ILLUSTRATION BY JAY BENDT



LAB LIFE

The path to purpose

Early-career researchers should persevere to find meaning in their work.

BY JACK LEEMING

When Melissa Craig realized that biological matter such as algae might slow the speed of powerful underwater landslides by releasing chemicals that help to glue sea-floor mounds together, she failed to share her colleagues' enthusiasm around the discovery. The group, based at Bangor University, UK, built miniature versions of those landslides in water tanks, changing the composition and amount of material that made up the mounds to model the sea floor at its most violent.

Craig's finding was unprecedented. "There had been nothing before her experiments that

found the same results," says ocean scientist Jaco Baas, Craig's supervisor at Bangor. But Craig, then a PhD student who was visiting from the University of Adelaide in Australia, couldn't see how her algal discovery would be of interest to anyone beyond her immediate academic circle. "I struggled to appreciate the impact of what I was doing," she says. So, in late February this year, Craig started working as a geologist with Oil Search, an energy company based in Papua New Guinea. She's hoping to conduct research that has clear practical applications.

Long before they embark on PhD programmes, potential scientists are told by teachers and the media that their work

will have a lasting impact. Whether it's by helping to cure diseases, build clean-energy infrastructure, or even provide food or water for communities affected by famine or drought, many early-career researchers hope that they can somehow make a difference. "We grow up thinking that we're going to solve the world's problems," says Florie Mar, a scientific-communications director at Genentech in South San Francisco, California. She earned her PhD in cancer biology in 2015 from the University of California, San Francisco.

Although researchers are often motivated by a sense of curiosity and the drive to explore how the world works, some also see science as a way of making the world a better place. ►

► It's widely acknowledged that the scientific endeavour works as an accumulation of small discoveries. "Your knowledge — however unpredictable and however useless it may appear — will be valuable," says Philip Blower, a cancer-imaging chemist at King's College London. But Mar and Craig say that they had hoped to obtain more tangible results during the course of their research. "I wanted something that I could measure," explains Mar. Craig reaches for a similar thought: "I wanted something visual — like seeing someone walk around with something that you know you made or helped with."

Both now feel that they're using their scientific skills to do meaningful work outside academia. Mar, for example, instructs doctors and pharmacists on the clinical uses of Genentech's drugs — information that she anticipates will go on to help patients.

The need to find deeper meaning in their work plagues many scientists, who can feel stuck in an often cut-throat system that's more concerned with bibliometrics than transferability. Many also say that they can't see how their research is contributing to society in a meaningful way. A 2016 report by professional-networking service LinkedIn (see go.nature.com/2i2srctc) notes that 41% of research professionals — compared with 37% across all sectors — say that they are mainly driven by purpose rather than by money or status.

Employers should sit up and take notice: the same report found that purpose-driven employees had greater job satisfaction and were more likely to stay at their company for at least three years. To help maintain an interest in their work, researchers can try a variety of ways to stay motivated. Some seek out laboratories that are engaged in highly translational research or that collaborate with partners in industry. Others look for research posts in industry. And more commercially focused scientists might even launch start-up companies to find the impact and value that they require.

IMPACT THROUGH INDUSTRY

For many early-career researchers, the sheer size of the scientific endeavour, as well as an increasing pressure to win grants and to publish results, can be discouraging. "When I talk to postdocs and PhD students, they often feel like cogs in a massive machine," says Jason Blackstock, a lecturer in public policy and engineering at University College London, who trained as a physicist. "There's tremendous pressure to publish in whatever the direction the UK research councils are funding in."

Dolores Del Prete, a postdoctoral researcher who studies the role of certain cells in the brain in psychiatric disorders at BioMed X, a contract research organization in Heidelberg, Germany, agrees that the system can feel overwhelming. During her first stint as a postdoc,



KEITH ARKINS

The creation of Almac Diagnostics in Craigavon, UK, required funding from outside academia.

in which she investigated a protein linked to Alzheimer's disease at Albert Einstein College of Medicine in New York, Del Prete decided that she needed to move into an environment that conducted more-applied research. "I was doing really basic research," she says. "I liked it, but I had this pressure that was all about publishing — to get more papers, to get more grants, to keep going. It was frustrating." She thinks that the publish-or-perish hamster wheel in which she became stuck will be a concept familiar to many researchers in academia. "It was research led by papers and grants, not papers and grants led by science," she says.

At BioMed X, Del Prete can see the results of her work more clearly. "In the future, industry can develop our drug," she explains.

For some, the ability to stave off frustration and to find meaning in their work can come from developing a side project. When Mar noticed her PhD research turning from an enjoyable endeavour into a three-year chore, she started to make videos that combine voice-overs and whiteboard illustrations to explain complex topics such as genetic variation, neuroscience and diabetes to the public. So far, she's posted more than 60 videos on YouTube as a way to engage with her audience, track interest, teach science — and find real-life, immediate impact. She has since launched Youreka Science, an independent science-communication company.

Mar says that the skills she acquired during her PhD, including critical thinking and communicating scientific ideas to the public, have supported her new career direction. "It's a way to utilize your scientific training," she says.

TRANSLATIONAL SCIENCE

Blackstock says that early-career scientists who want their research to have wider societal impact should seek to work and study at

institutions that nurture this drive through programmes that focus on transferability. "If you really want to change the world, just learning the technical stuff still leaves you trying to figure out how anything you've learned matters," he says. Instead, "Find programmes that have really strong practical levels of engagement. Or, at the very least, programmes whose projects have real-world partners" that can teach students to apply their findings in a practical way outside academia. As an example, Blackstock cites University College London's 'How to Change the World' training programme, which he directs, and which pairs engineering students with representatives from industry and government to examine issues such as water quality and public transport. Partners have included the UK Department of Transport and London-based engineering company Atkins.

He also suggests that students should find PhD programmes that collaborate directly with industry. Germany's Fraunhofer Society, for example, receives 70% of its research funding from contracts with external partners and operates institutes that focus on topics such as lasers and wood technology. Developing a device or procedure that people need is an integral component of the research programme rather than a bonus.

After Atma Ivancevic completed her PhD in bioinformatics at the University of Adelaide, she realized that she would need to take care in selecting a lab for her postdoctoral work. Ivancevic had studied mobile elements in DNA — sequences that move around the genome across generations to drive evolution — and found it enjoyable, but says that its transferability was not obvious to her at the time. "It was hard to see how it could be applicable to something these days," she says. "The sorts of effects we're talking about take millions

of years to come to fruition.”

She knew that she had found the right lab when, during an interview, a potential supervisor tilted his computer screen towards her. “He showed me four or five e-mails he’d got that day from families — not from other scientists,” says Ivancevic. One asked about his research on the genetics of severe epilepsy in women and whether there had been any recent advances. “It didn’t matter if he published a paper that year or not,” she says. “He still would have answered those e-mails. That’s real-life impact right there.”

Ivancevic thinks that luck, as well as design, might play a part in determining the impact of scientists’ work. “Maybe they just haven’t found out how it is applicable yet.” Blower also believes in scientific serendipity, and therefore advocates for research that doesn’t always set out to solve a specific problem or address a specific issue. “You turn over loads of stones and, with most of them, there’s nothing underneath, but occasionally there’s something. If you don’t turn over the stone,” he says, “you don’t find the thing.”

One such stone revealed the gene-editing technology CRISPR. Rachel Haurwitz did her PhD and worked as a postdoc in Jennifer Doudna’s lab at the University of California, Berkeley — one of the birthplaces of CRISPR. Haurwitz, now chief executive of Caribou Biosciences in Berkeley, which aims to commercialize the technology, sees the rise of CRISPR as evidence to support the continued funding of basic research, alongside more translational work. “I think this story further cements the tremendous value and need for investing in basic research,” she says. “To pretend that we know exactly where to go to discover or invent the next big thing is incredibly naive.”

She suggests that scientists who want to see the impact of their work should seek out labs and companies that do translational research. “Actively find a way to put yourself there,” she says. “There are some labs in academia that are closer to that boundary, and there are lots of companies in industry who use life science and technologies to try to solve a problem.”

REAL-WORLD MOTIVATION

For some researchers, launching their own business can provide the meaning that they seek. In 2000, Paul Harkin, a molecular oncologist at Queen’s University Belfast, began to realize that to extend his work on the gene *BRCA1*, which is implicated in hereditary breast cancer, he had to move away from academia.

Harkin had recognized that preserved samples of tumours stored at labs and

hospitals worldwide would be an invaluable source of data that links genetic information with patients’ outcomes, if clinicians had the tools to reliably extract partially degraded RNA from the tissue. But he was unable to launch the project from his lab at Queen’s. “I needed to bring in substantial funding and additional expertise to get to commercial application,” explains Harkin.

So, in 2004, he co-launched a company — now known as Almac Diagnostics and based in Craigavon in Northern Ireland, UK — to take his work to market. “I’ve never been disillusioned,” he says, “but I was very pragmatic about what could be achieved in an academic environment.”

The company’s focus has since pivoted to providing clinical-trials support to the pharmaceutical industry. And Harkin notes that at least one of the drugs that it has worked on has been marketed in the United States.

Although he draws satisfaction from knowing that the company he built is directly involved in getting medicines to patients, Harkin highlights the positive effect that Almac Diagnostics has had on the scientific-employment landscape of Northern Ireland. He estimates that around 50% of Almac Diagnostics’ employees hold PhDs; and its parent company, the Almac Group, employs more than 3,000 people in the province. “There are now alternatives in the scientific arena in Northern Ireland — it’s not just jobs in academia,” he says.

Harkin thinks that early-career researchers who want to make an impact should seriously consider accepting a position in industry. “Young scientists coming through don’t understand the potential in industry,” he says. “You may not own a project in its entirety, but you’re part of that team that gets something into the clinic.”

Yet many scientists maintain that curiosity is enough to justify investigating a research question. Baas’s interest in sea-floor deposits is driven by a wonder at how the world works. “What motivates me is discovering things, really,” he says. “I have questions in my head all the time; I want to find answers to those questions. Research is the ideal vehicle to do that. My work is my hobby.”

Ivancevic is set to begin another postdoc in August. She says that even if she had left academia, she would have stayed up to date with research in her field, and understands the drive of curiosity. “I can see how it can consume you,” she says. “You just want to find out why.”

Craig also expects to keep track of her academic field. “It’s almost like a hobby — it’s so cool and significant to the geoscience community,” she says. “But I’m still drawn to other pursuits that apply my science.” ■

Jack Leeming is the editor of *Naturejobs*.

INTERNATIONAL STUDENTS

A shift in interest

A report that gauges the preferred destinations of prospective students from around the world suggests that the United States and the United Kingdom are losing their appeal for students from some regions. ‘Applicant Survey 2018: What Drives an International Student Today?’ — conducted by London-based educational-marketing group Quacquarelli Symonds during the 2016–17 academic year — finds that more students than before are aiming for Canada, Australia or elsewhere. Overall, 48% of the 16,560 students surveyed listed the United States as one of their preferred destinations. The United Kingdom came in second at 42%, followed by Canada at 34%, and Australia and Germany at 28% each. The survey found that Canada had risen in popularity with prospective students from all regions, and had replaced the United Kingdom as the second most popular destination for respondents from Latin America and the Middle East and Africa. The United States had declined in popularity in some countries in Africa and the Middle East. The report speculates that the election of Donald Trump as US president and the UK Brexit vote might have influenced respondents’ indications of interest.

UNIVERSITIES

Fewer women at the top

Female leadership at 200 of the top-ranked universities worldwide fell this year to 17%, according to a report. Just 34 of the universities named in the 2018 *Times Higher Education* World University Rankings have female presidents, compared with 36 last year. Among the listing’s highest-ranked institutions across 27 nations are the University of Oxford, UK; Harvard University in Cambridge, Massachusetts; Imperial College London; the University of Pennsylvania in Philadelphia; and the University of California, Berkeley. The rankings consider research, teaching and international outlook among other factors. In Sweden, 4 of the 6 institutions that made the list are led by women. The United States has 11 female-led universities in the rankings, the report’s highest number. Janet Metcalfe, head of Vitae, a UK-based advocacy group for researchers, expressed concern at the figures. “More women in leadership positions provides positive role models for female academics,” says Metcalfe, “and can encourage better gender balance and diversity at all levels.”

REQUIEM

Return journey.

BY CHRISTINE LUCAS

Is it an invasion if it's only one alien? Some media did call it so, even if this alien didn't come with guns a-blazing. Was he even male? No one could tell, but from the start the media defaulted to their perceived gender of an assumed conqueror. He came down from the sky in his craft of light — his pod, or capsule, or perhaps cocoon. With a smooth, effortless descent he landed in the Sahara Desert, on a stretch of dirt and yellow shards where millennia ago, before Ramses and Alexander, intense heat had turned sand to glass. And then the web erupted with triumphant screeches about the old gods returning to lead mankind to ascension.

Only he didn't.

He strolled through the streets of Cairo with the familiarity of someone born there. He didn't actually walk — he *glided* a couple of centimetres above the ground, clad in his ankle-long, shimmering garment that could be a robe or a kaftan, and browsed the stalls and booths of Khan el-Khalili just like another tourist. His dark, lidless eyes scanned fabrics and glassware, and perhaps lingered a little too long on the miniature cat-shaped statues displayed on the *souq* stalls before seeking the wide-eyed, living cats that had inspired them.

Long fingers — an artist's fingers, some argued, too delicate for warfare — brushed against spice racks, dried fruit and flatbread amid other, non-edible wares. Then he stopped before a street-food cart and its selection of grilled meats: *kofta*, *kebab*, *shawarma*. Under the flabbergasted gaze of the petrified vendor, he broke off the tiniest piece of lamb meat and brought it to his mouth. Something that could be a smile lit up the expressionless face, and something that could be words left his thin lips. Then he stood there for a moment that stretched on, awaiting a response that never came, save from the vendor's white-knuckled grip around his *nazar* amulet.

And then he vanished.

He appeared again sampling noodle soup in Hong Kong, then a shot of espresso in Rome, then in Peru and Mumbai and Singapore, and several other places around the world. He ventured through local markets, sampled their food and, some believed, tried to engage the locals in conversation. The media crews flocked behind him with every possible recording device, and they all failed. His garments emitted a force field that



rendered all electronics useless after a certain radius. Electronics *and* bullets. Because when the lunatics came — how could they not? — assigning human stupidity to one deity or another, no bullet or bomb affected his stroll through the eateries of Earth.

And while theories on his origin and purpose raged, while the cooks and vendors he visited became celebrities overnight, his glow seemed to diminish with every new visit. Something did affect him.

And then he was gone. No one saw him for two weeks, and many assumed he'd left. Or died.

Until he appeared again, this time on a backwater island of an indebted country in the Mediterranean. No markets this time, only an old fisherman cooking fish fresh from his nets, and the alien perched upon a boulder a few metres away. The lone drone that caught this exchange recorded the alien's lips moving, and the fisherman, grey and withered like his boat, shaking his head.

"Ithaca? No Ithaca," he told the alien in broken English. A tourist is a tourist — what else would he speak? Then the old man pointed westwards. "This, Aegean. Ithaca, other sea. Ionian sea."

The old man poked the pile of embers and ash where something was roasting. He retrieved a package wrapped in aluminium foil and unfolded it. He held up a sardine enclosed in a thick salt crust.

"Here. Try, *phile*."

Phile. Friend. Did the alien eyes widen at the word?

➔ NATURE.COM

Follow Futures:

🐦 @NatureFutures

📘 go.nature.com/mtoodm

The fisherman leaned closer. "Good. Old recipe. My grandpa's grandpa's. Older.

Back to Odysseus."

"Odysseus," said the alien and shattered the crust to reach the fish meat inside. He took one bite. "Good."

As the old man reached for a second sardine, a woman emerged from the little house behind them, with the whitewashed walls and the blue windows. The drones that had swarmed to the area ran her face through every possible database: a retired schoolteacher, the fisherman's wife, who once taught an ancient, useless language to bored teenagers. Just another nobody in the long string of nobodies the alien had approached. She came bearing gifts of alcohol, and the alien mumbled something to her.

A moment of wide-eyed silence, then a slow shake of her head. "No. Not *oinos kekramenos*. We don't drink that anymore. Here. Try that. Ouzo. Good. Better than ambrosia."

If doubts slowed the alien's initial gulps of the star-anise-smelling liquor, they were washed down with the second bottle and the sardines that accompanied it. Then other dishes came out: olives and grilled peppers and hard bread softened with olive oil and topped with goat cheese.

"Not your first time here?" she asked the alien.

He licked the oil from his fingers and nodded.

"Long ago?"

Another nod.

"Why now?"

A long stare. He held out his open palm. The visage of an ancient copper *drachma* flickered on it. His palm closed a little too slow, as if in pain.

She sighed, and nodded, and offered him more ouzo. "One more for the journey. And for the Ferryman."

Beneath the crescent Moon, three old-timers ate and drank their night away, two of them counting their lifespans in decades, the third in millennia. When dawn came, they were gone. Only then did humankind realize how much the alien craft had looked like a coffin. But they never found it to confirm it, nor did they find the old couple, and soon they were all forgotten.

Until another ship came. ■

Christine Lucas is a former Air Force officer from Greece. Her work has appeared in several online and print publications, including *Daily Science Fiction*, *Space and Time Magazine* and *Cast of Wonders*. She was a finalist for the 2017 WSFA award.

ILLUSTRATION BY JACEY